

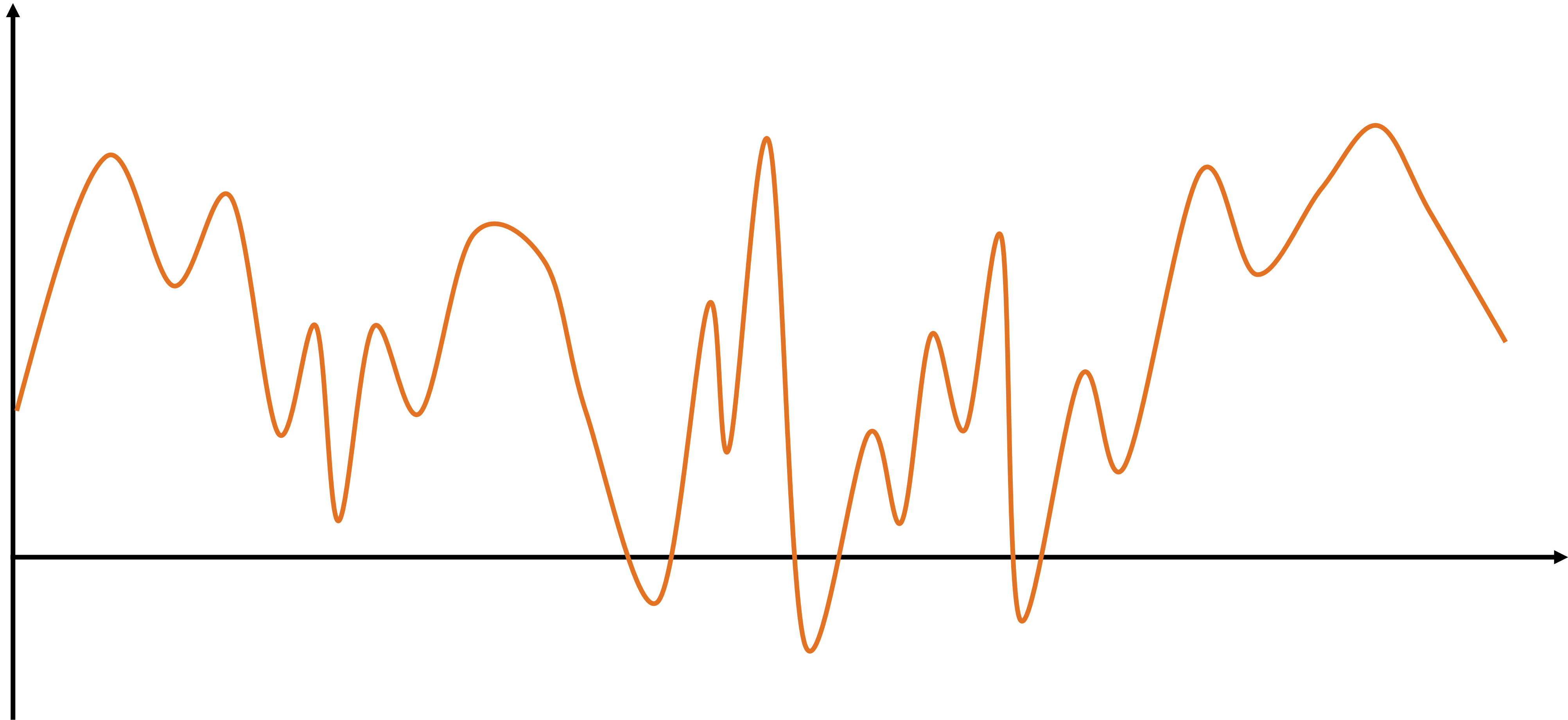
Digitalization

Mirco Schönfeld
University of Bayreuth

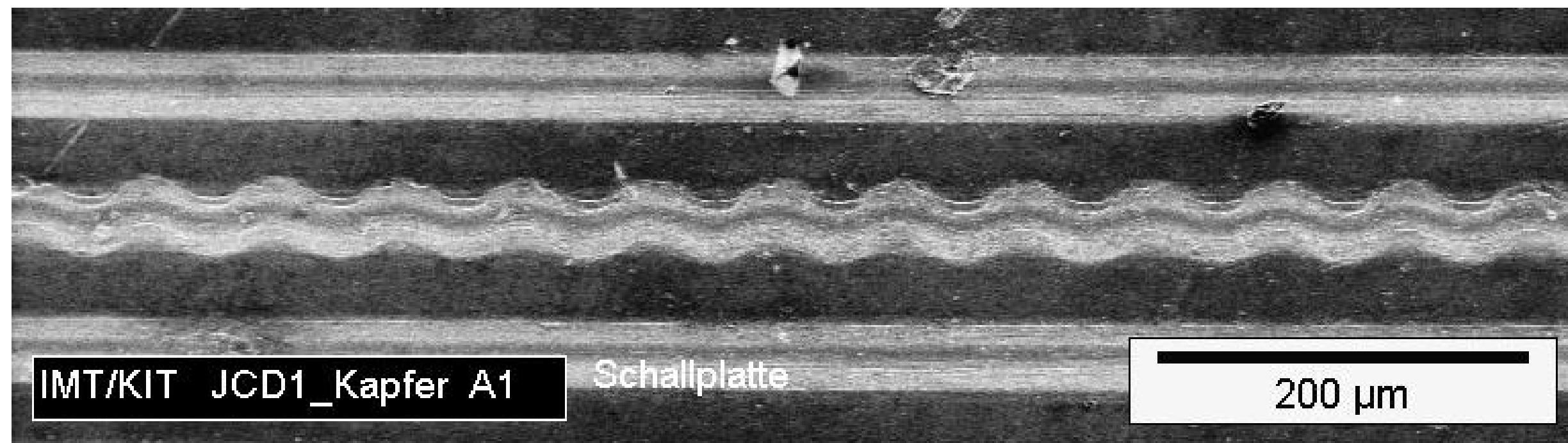
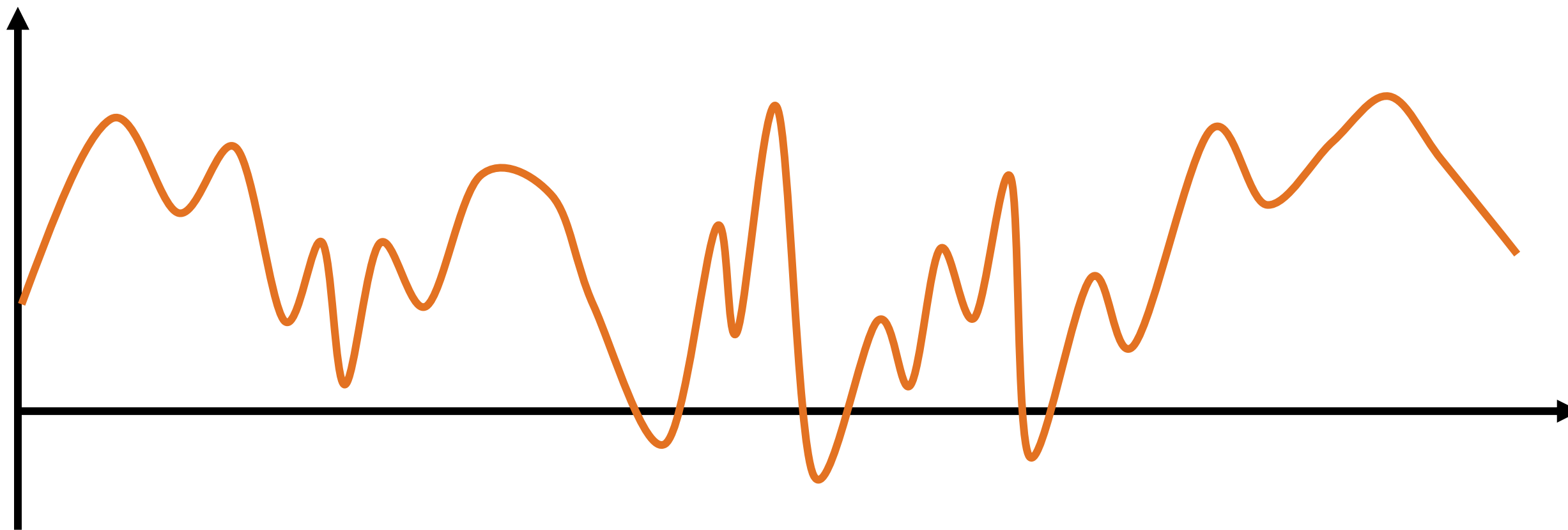
mirco.schoenfeld@uni-bayreuth.de
[@TWIyY29](#)

From Analog to Digital

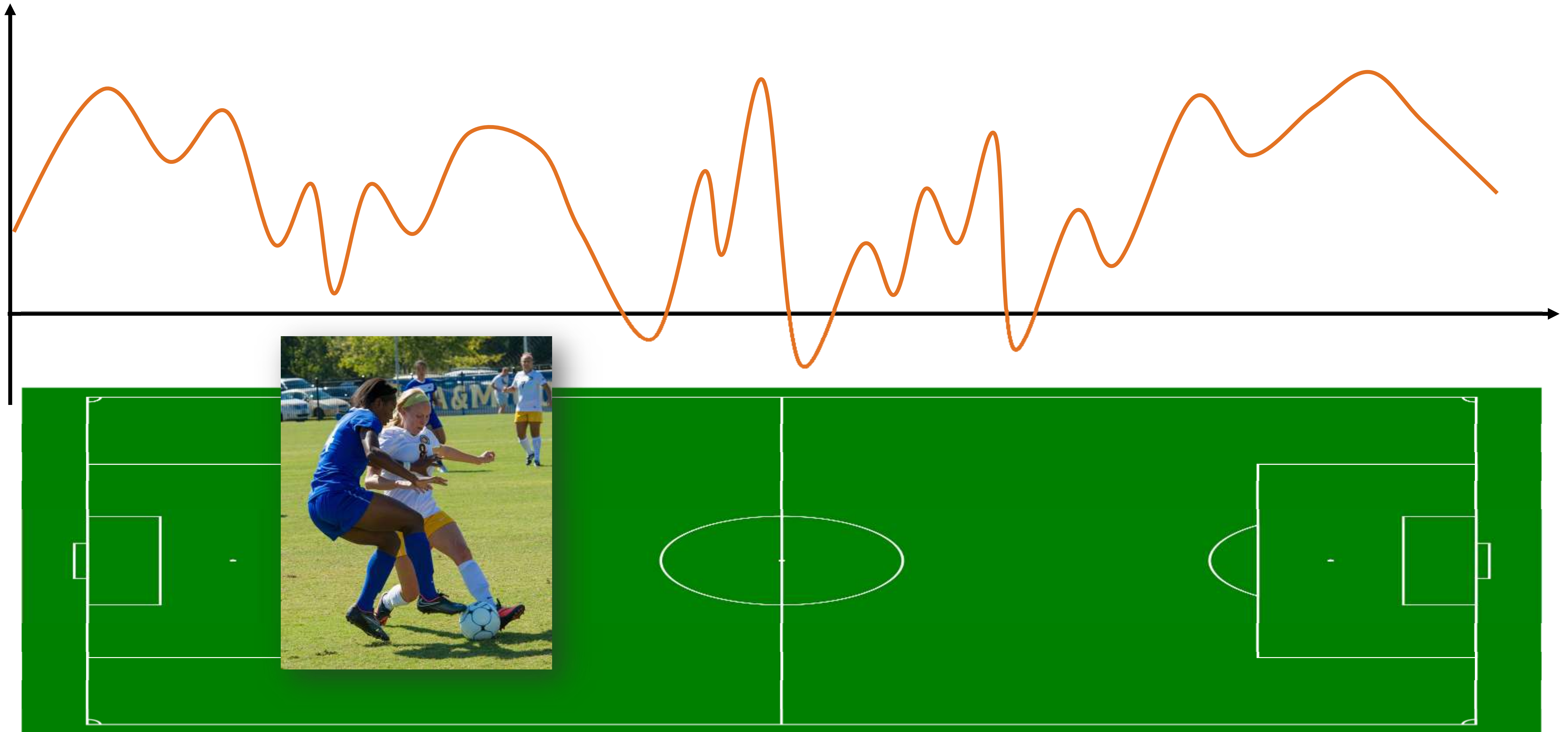
An Analogous Signal



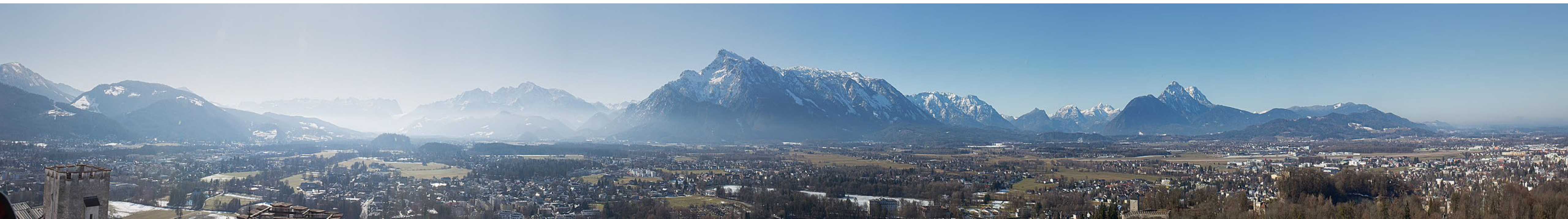
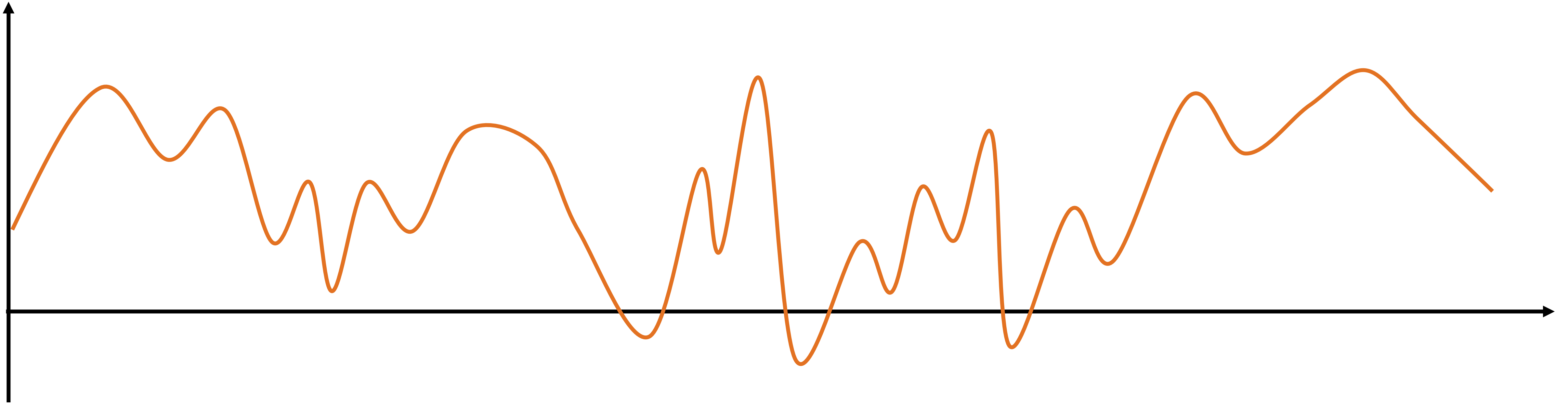
An Analogous Signal



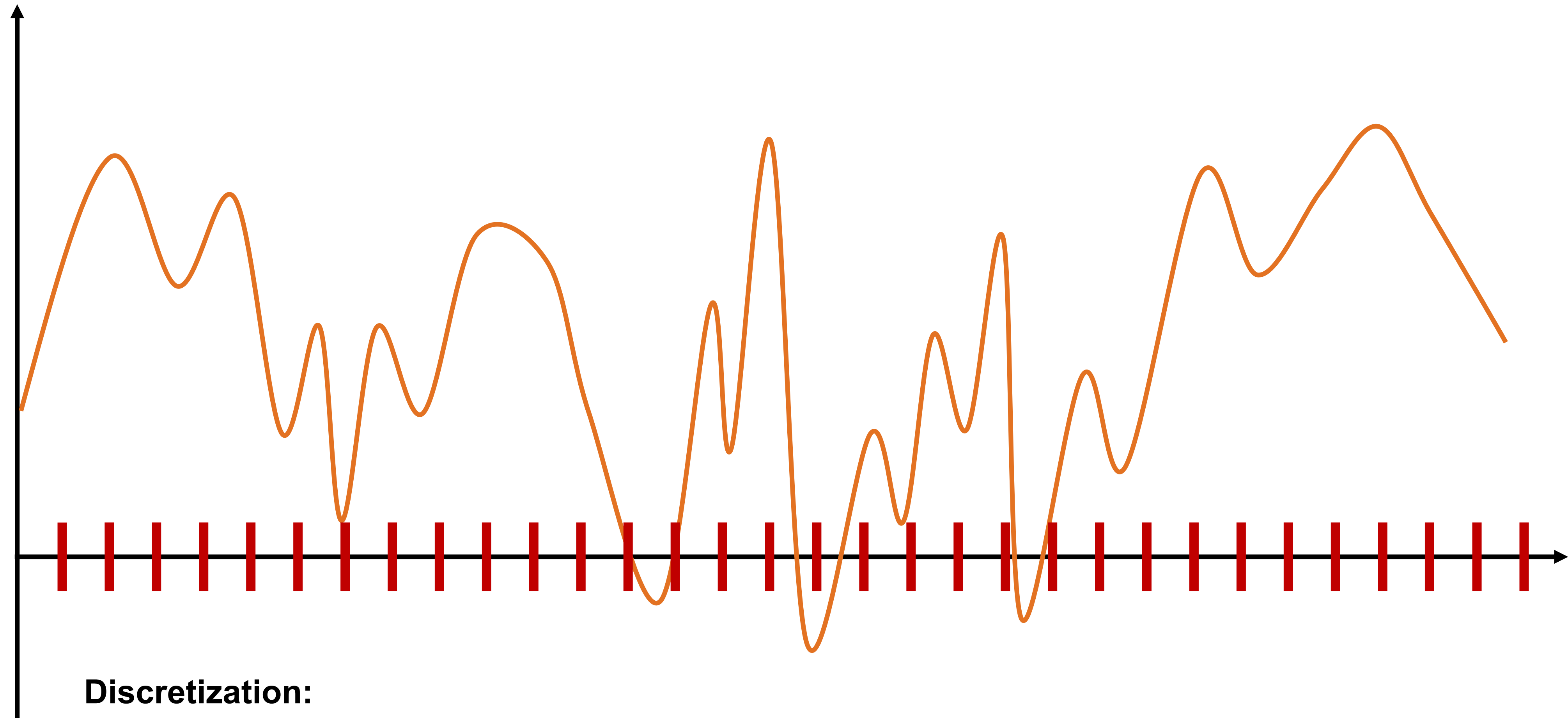
An Analogous Signal



An Analogous Signal

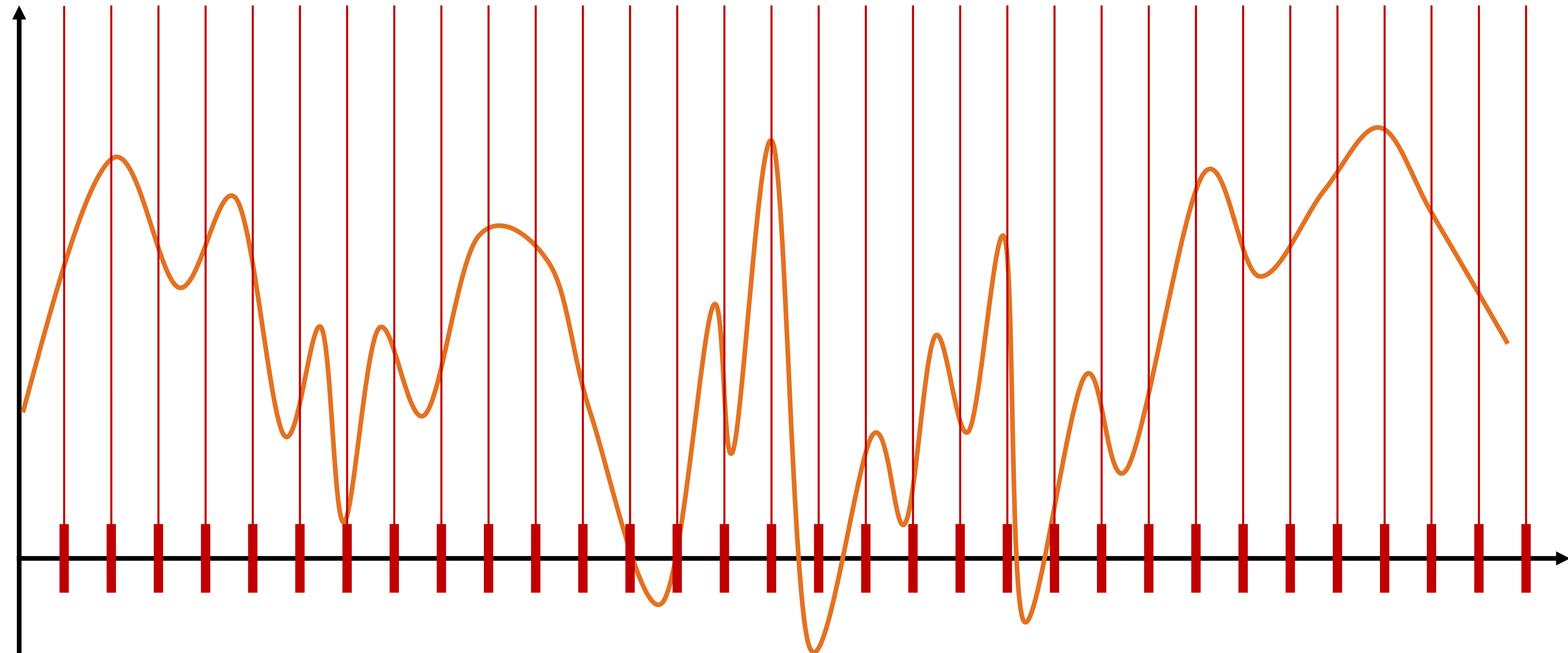


Discretization

**Discretization:**

Fixing a grid of measuring points is defined on the axis over which the signal changes (time, space, ...)

Discretization

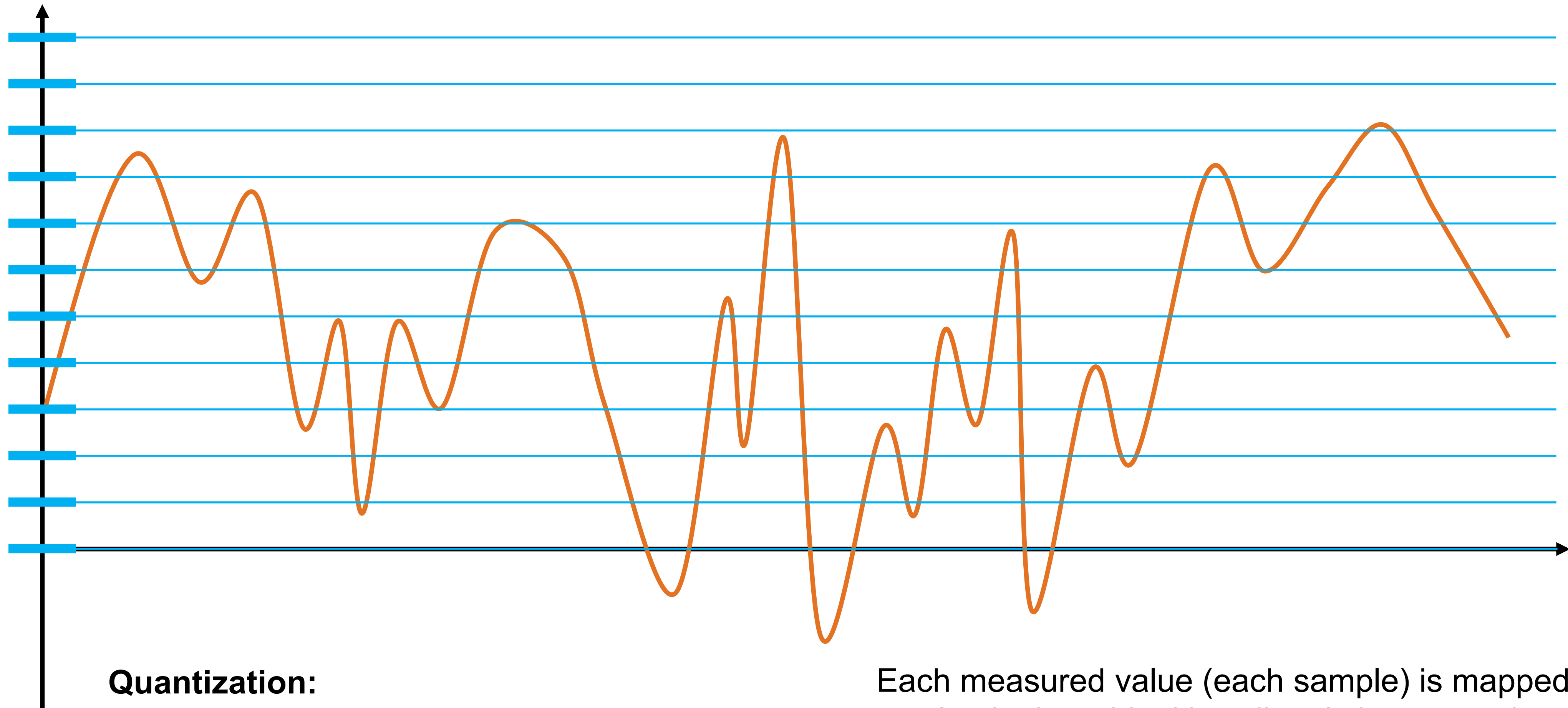
**Discretization:**

Fixing a grid of measuring points is defined on the axis over which the signal changes (time, space, ...)

Sampling:

Determining the current value of the signal (*sample*) for each measuring point.

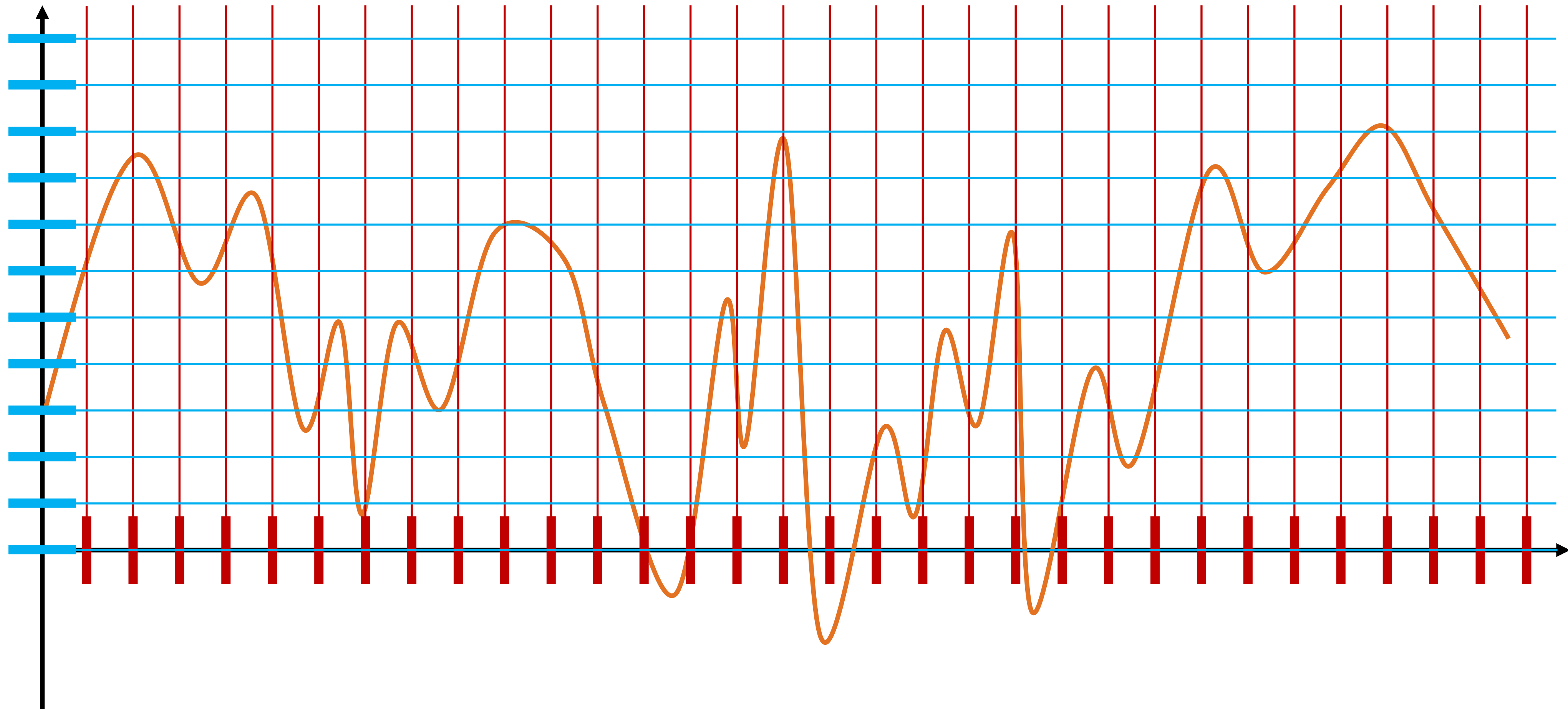
Quantization

**Quantization:**

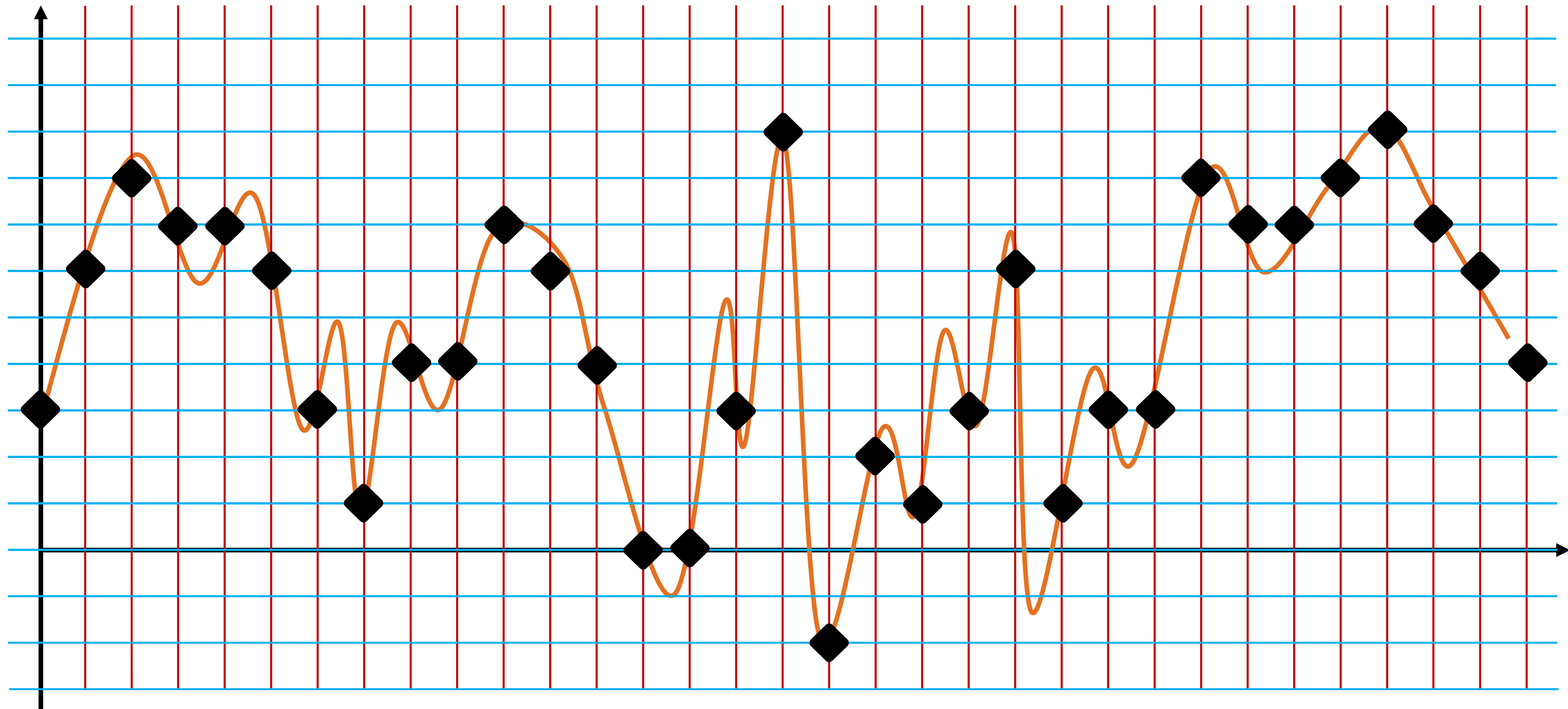
Representation of the measured values in a fixed whole-numbered value grid

Each measured value (each sample) is mapped to a value in the grid, either directly by measuring instruments or by calculation (e.g. rounding) from analog measurements.

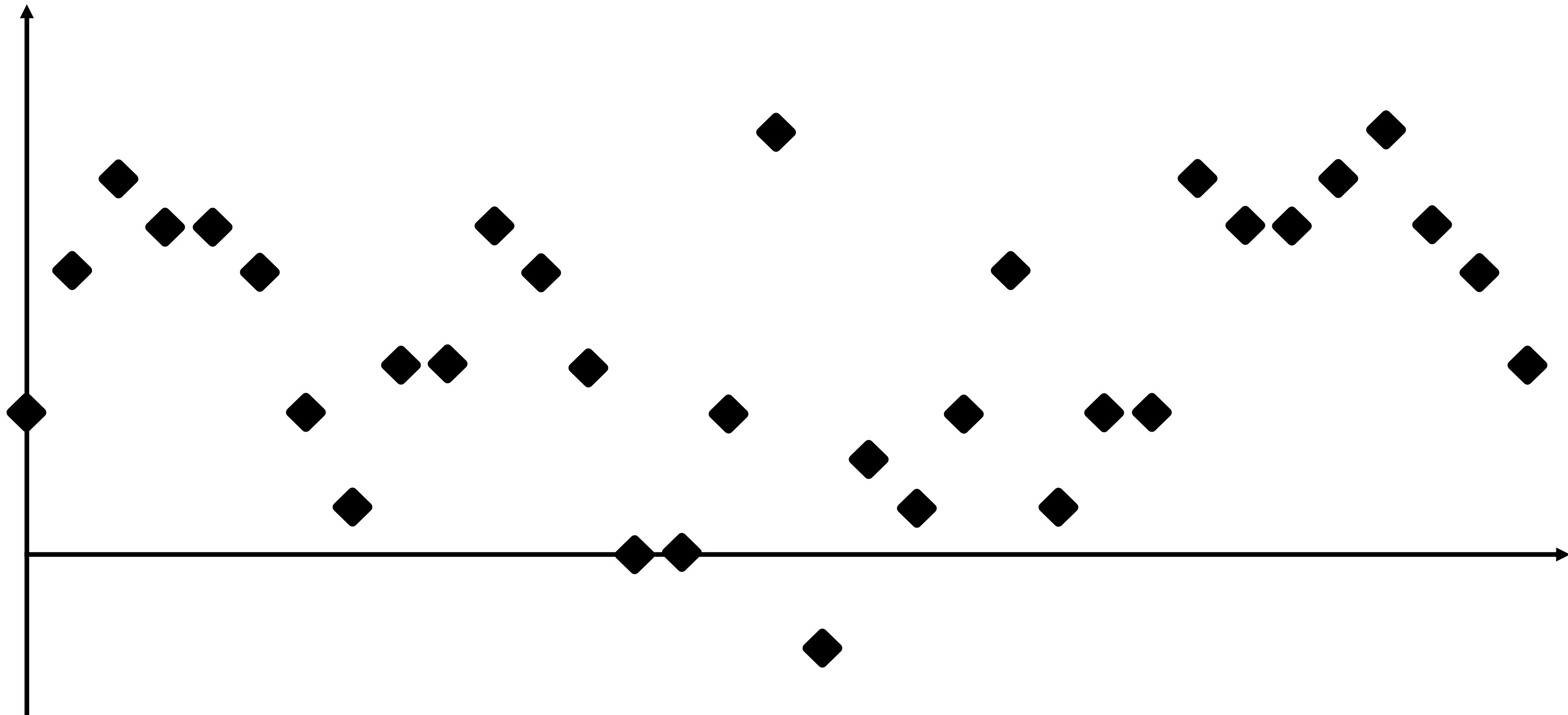
Discretization + Quantization = Digitization



Discretization + Quantization = Digitization



A Digital Signal



Bits and Bytes

Bits: smallest unit of storage, atomic, either 0 or 1

1 Bit is too small to be of much use

8 Bits together: 1 Byte

1 Byte can store:

- 1 number between 0 and 255 or
- 1 character

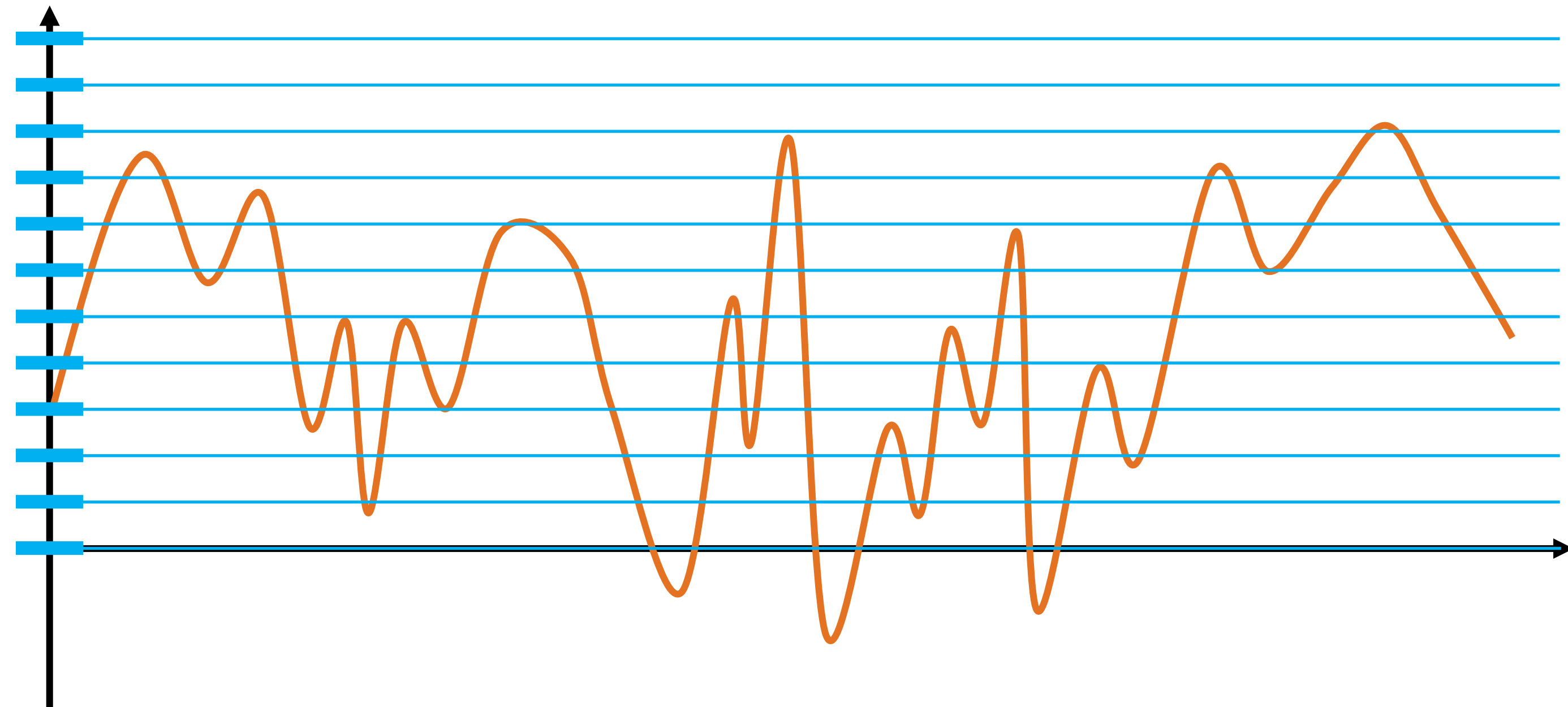


Number of bits	Patterns
1	0 1
2	00 01 10 11
3	000 001 010 011 100 101 110 111

Bits and Bytes for Signals

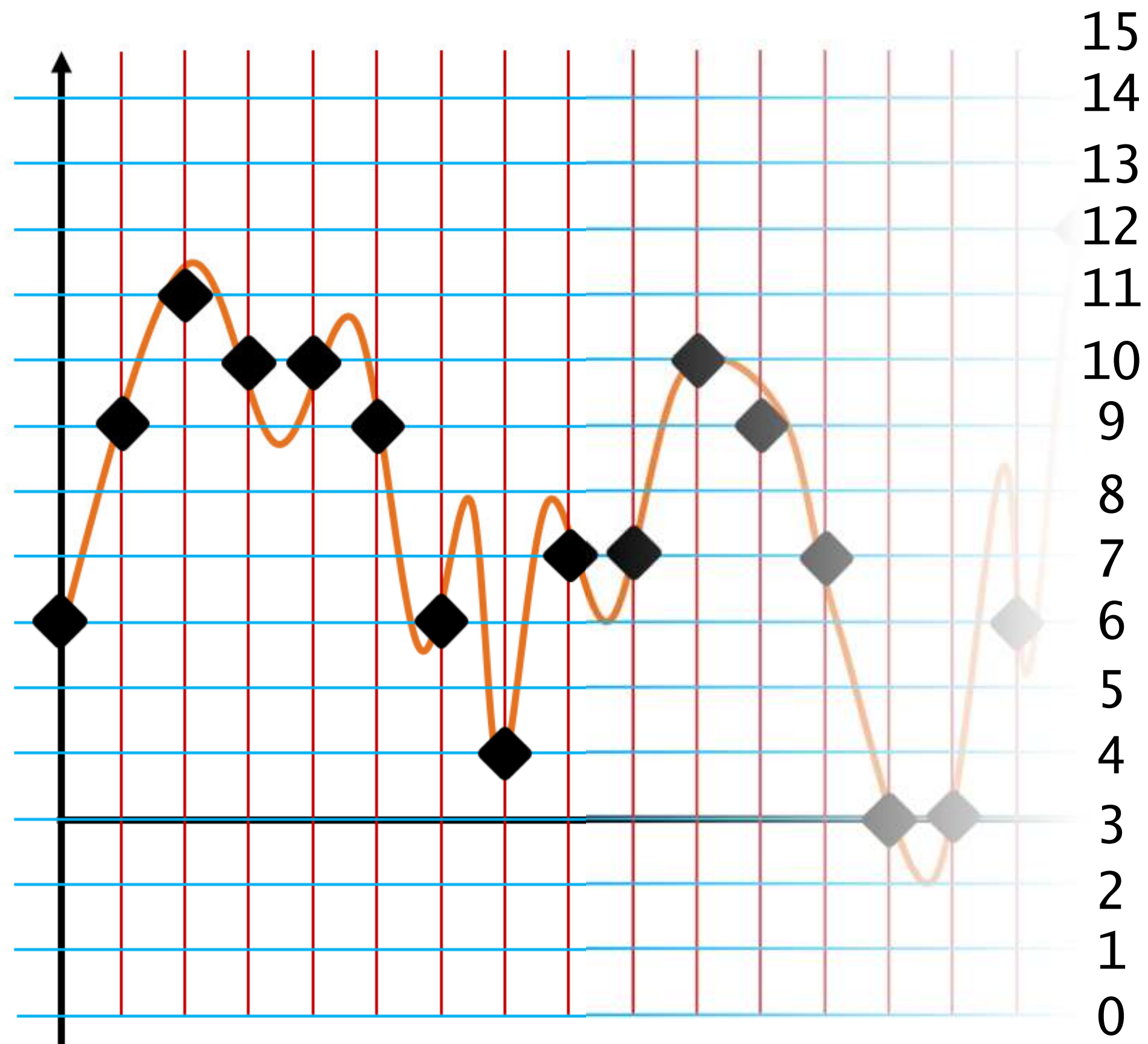
The number of steps in the quantization grid decides upon the required number of bits:

- 1 bit - 2 patterns
- 2 bits - 4
- 3 bits - 8
- 4 bits - 16
- 5 bits - 32
- 6 bits - 64
- 7 bits - 128
- 8 bits - 256 - one byte



n bits yield
 2^n different patterns

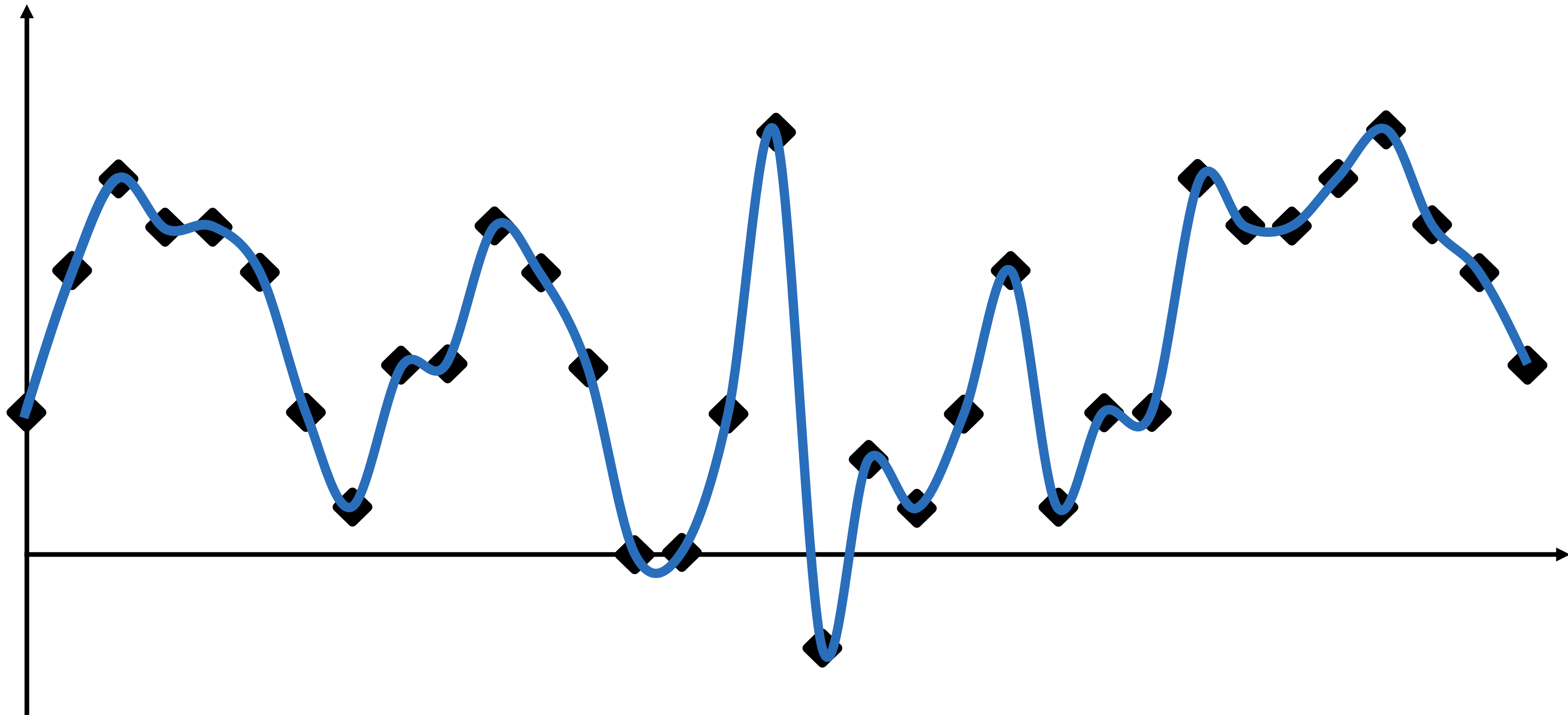
A Digital Signal



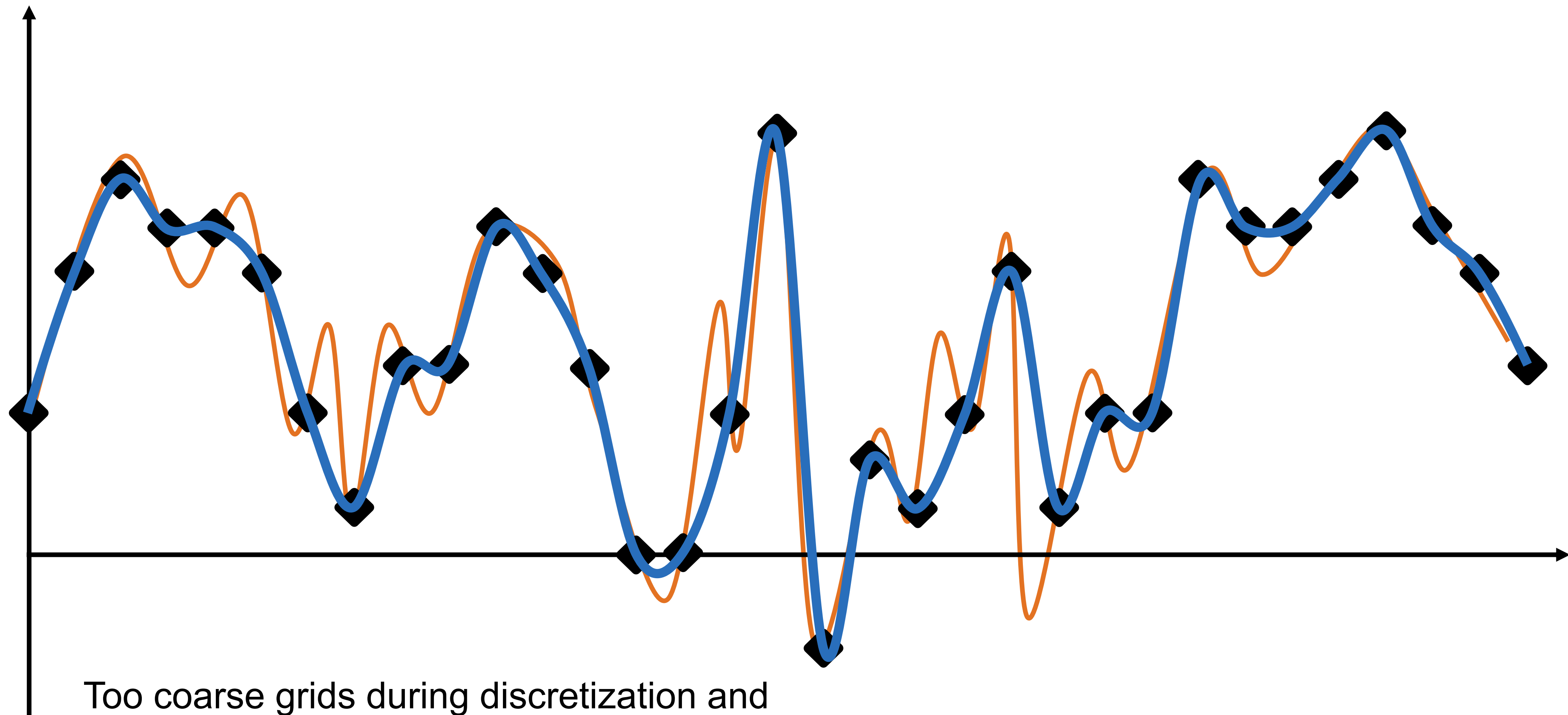
Time	Value	Pattern
1	6	0110
2	9	1001
3	11	1011
...		

The Signal:
0110 1001 1011 ...

Recovering an Analogous Signal



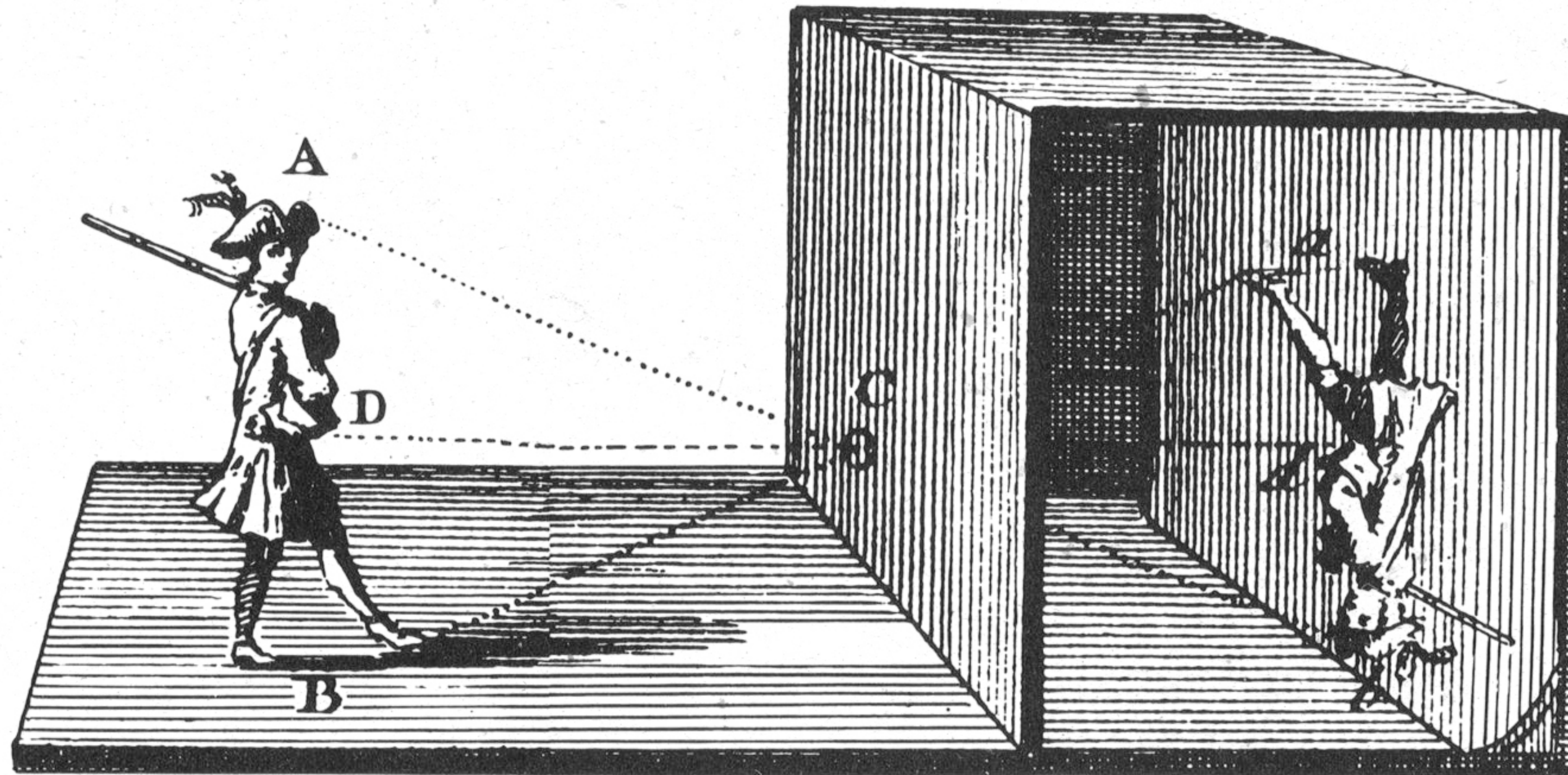
Digitizing Errors



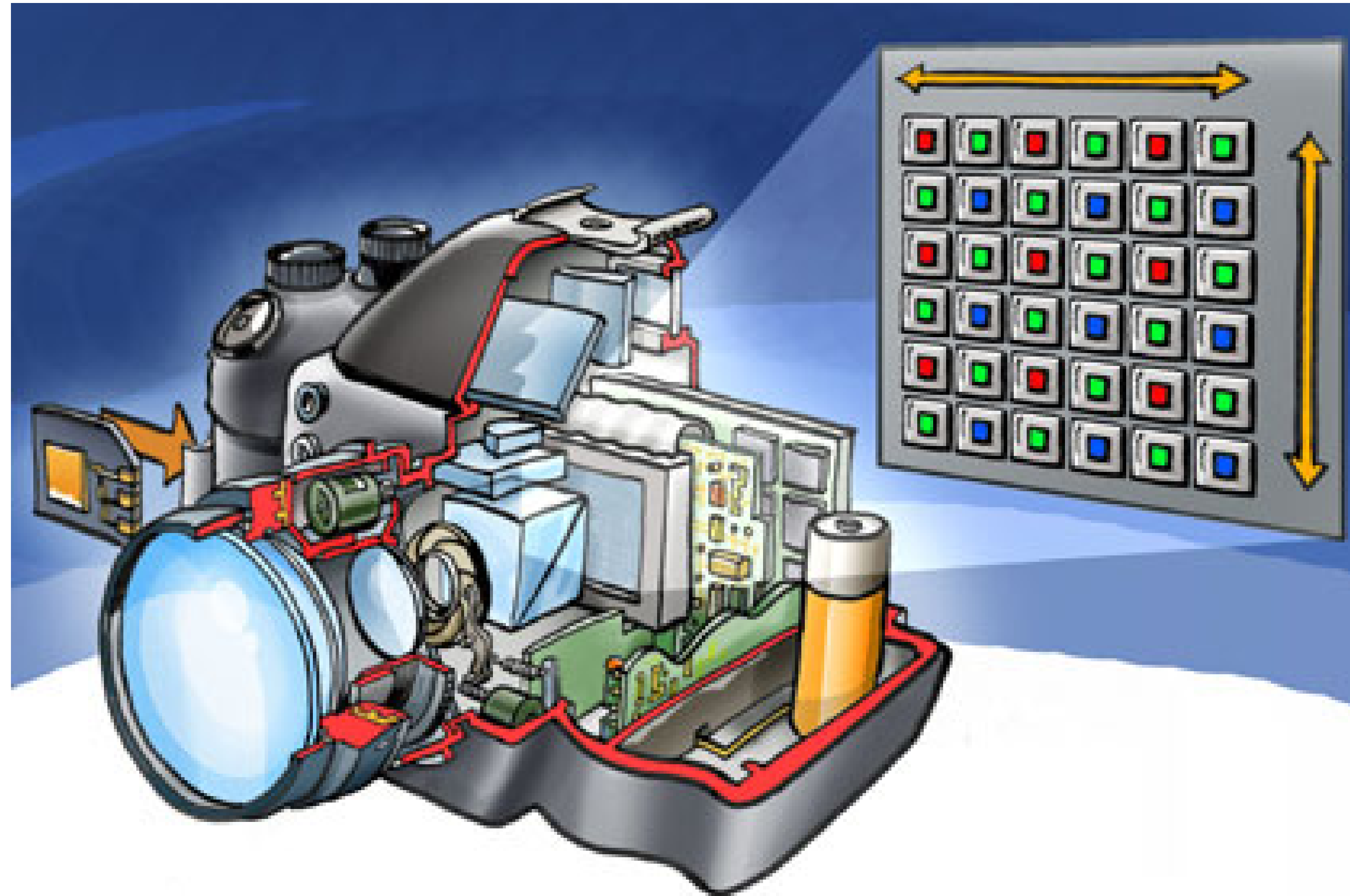
Too coarse grids during discretization and quantization result in **digitization errors**.

Digitizing Images

Camera Obscura



Digital Camera

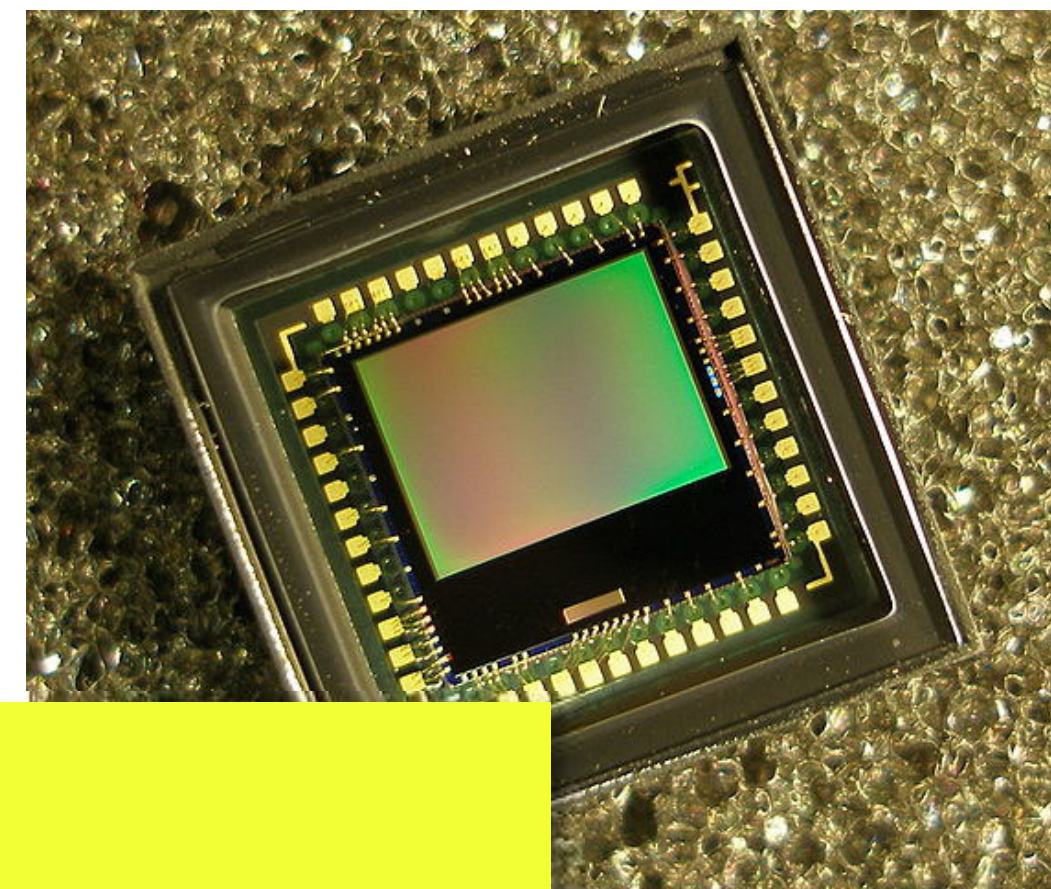
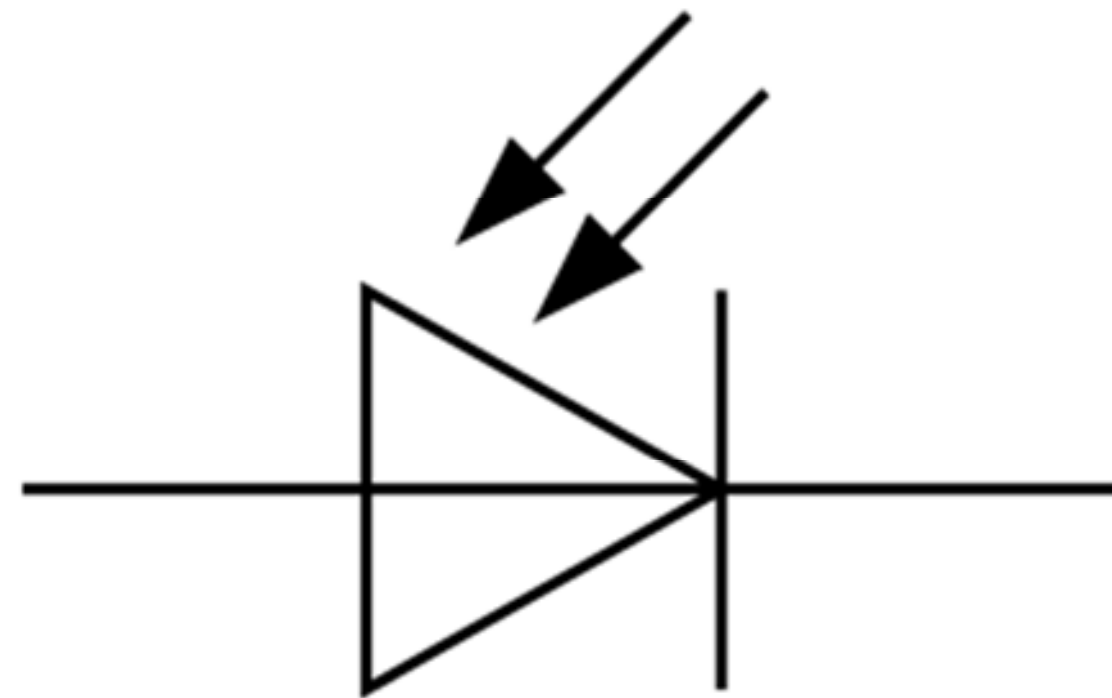


Analog-to-Digital Conversion of Light

Basically, a photodiode on a sensor translates luminance to a voltage and thereby reports the luminance of *one pixel*.

On approx. 12-35mm², current CMOS sensors have 12-108MP, i.e. 12-108 *Million Pixels*

A Photodiode takes up
around 0,7µm in a
current camera phone



CMOS Sensor

Where does the
color come from?

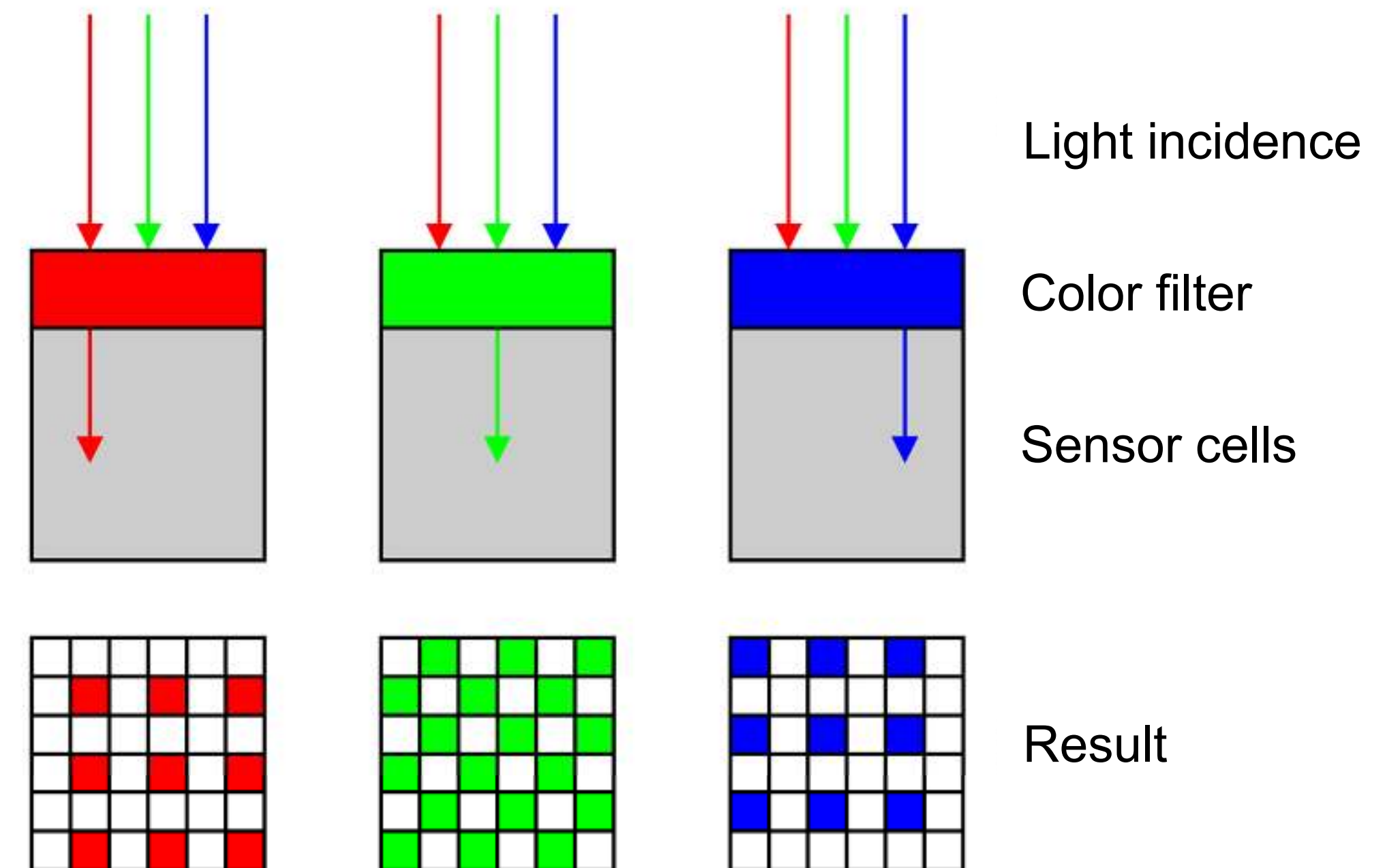
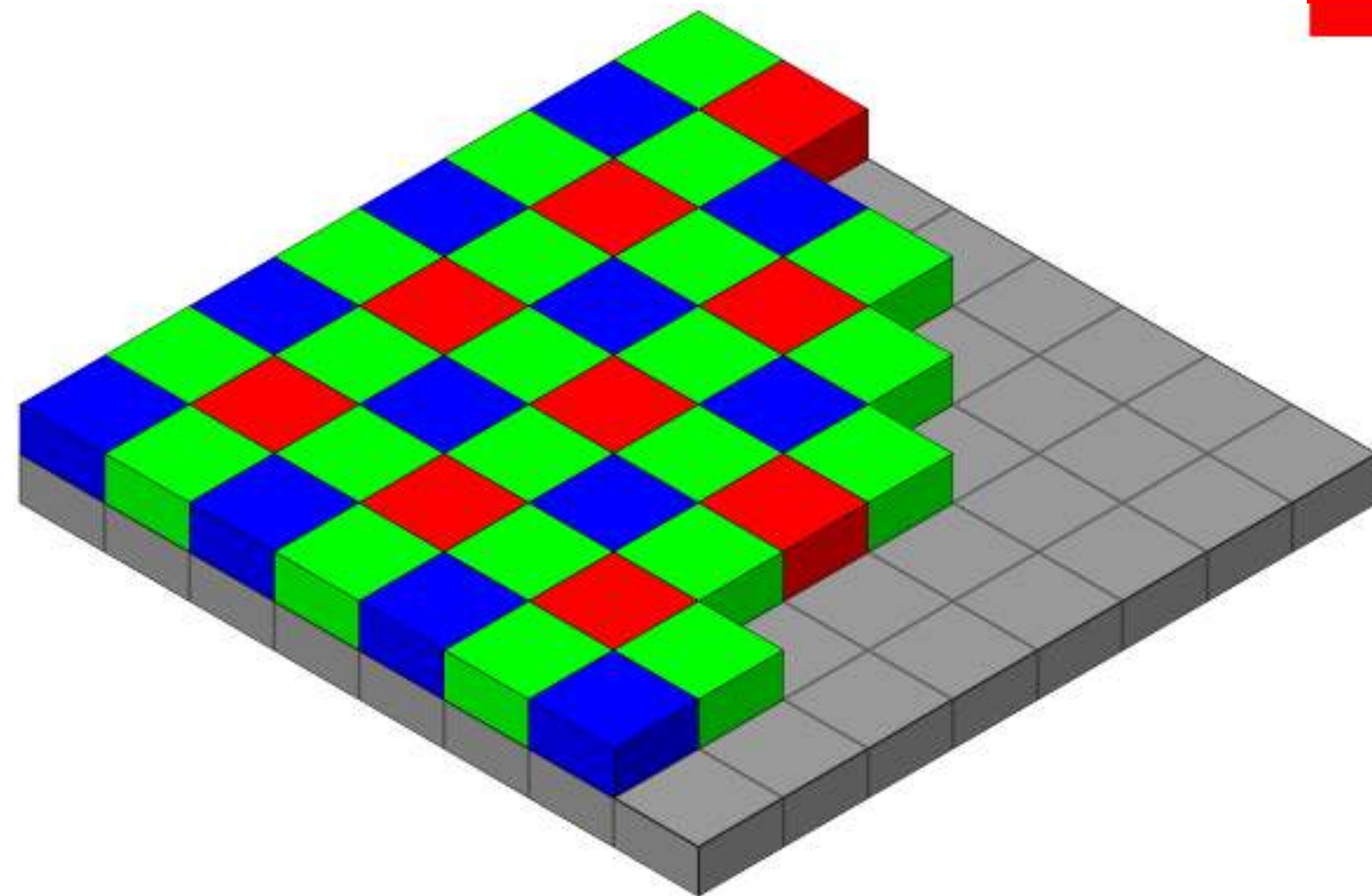
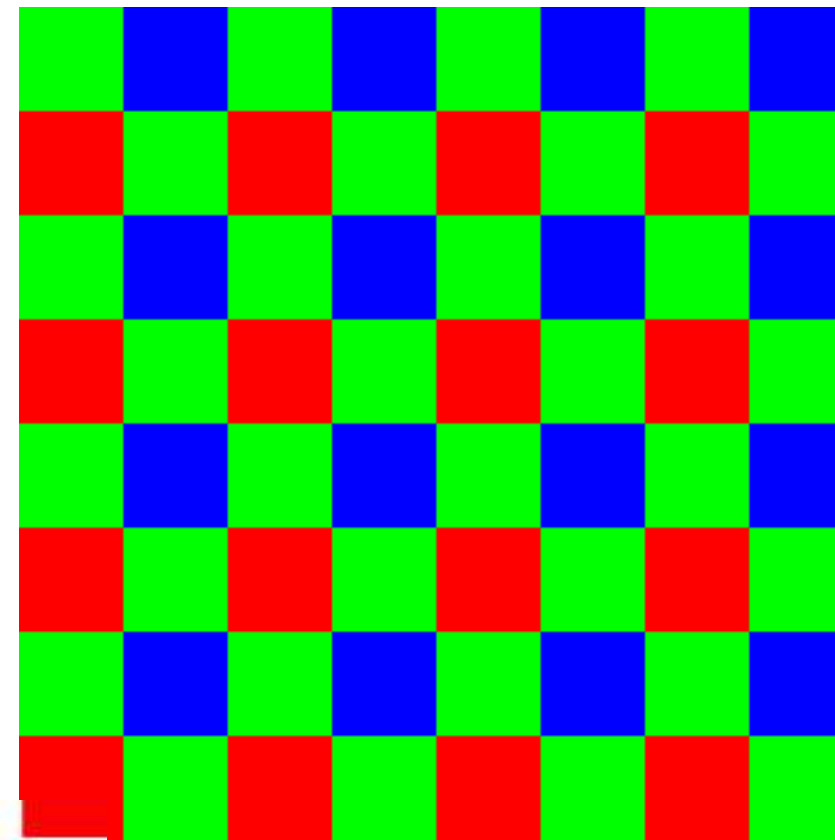
Bayer Sensors and Filters

Arrangements of RGB color filters on sensor arrays containing

50% green

25% red

25% blue filtered cells



Explanation:

Green component in gray tones makes the greatest contribution to the human eye's perception of brightness and thus also to the perception of contrast and sharpness.

72% of the brightness and contrast perception of gray tones is caused by their green component, whereas red contributes only 21% and blue only 7%.

Sensor Architectures

Earlier: CCD (*Charged Coupled Device*) sensors,
Mostly replaced by CMOS (*complementary MOS* or *Active-Pixel-*) sensors.

Both chip designs have their dis-/advantages



Disadvantage of CCD: Blooming



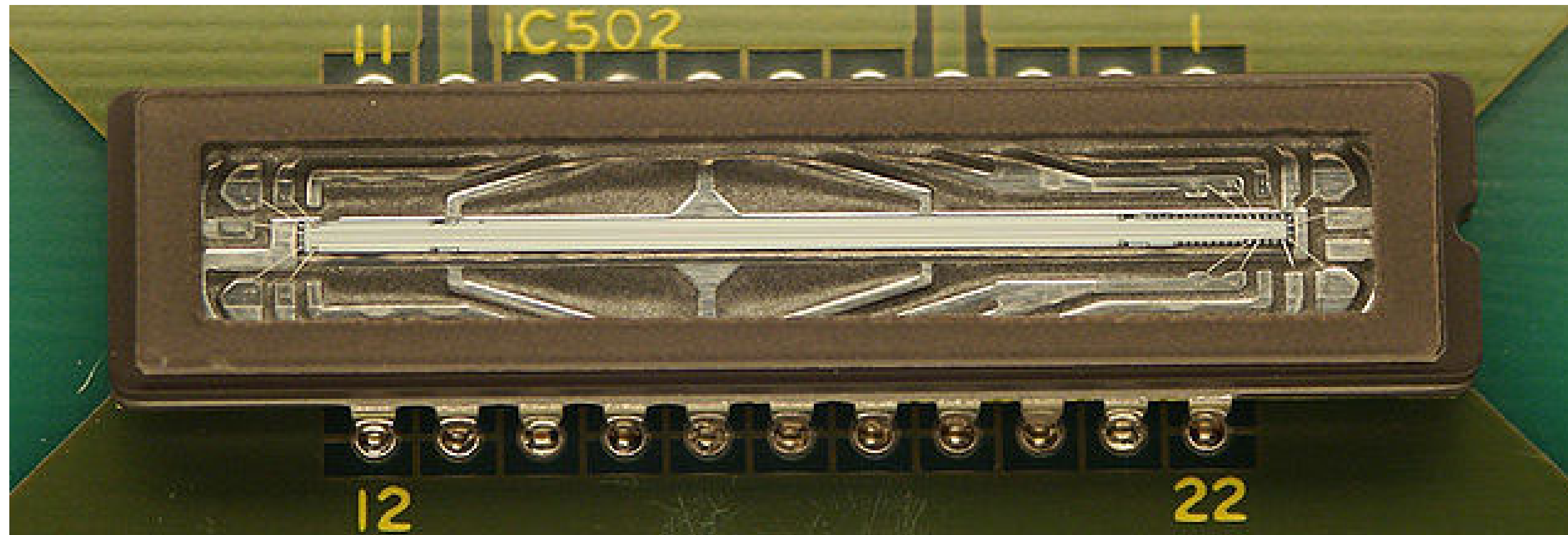
Disadvantage of CMOS: Rolling Shutter

Scanners

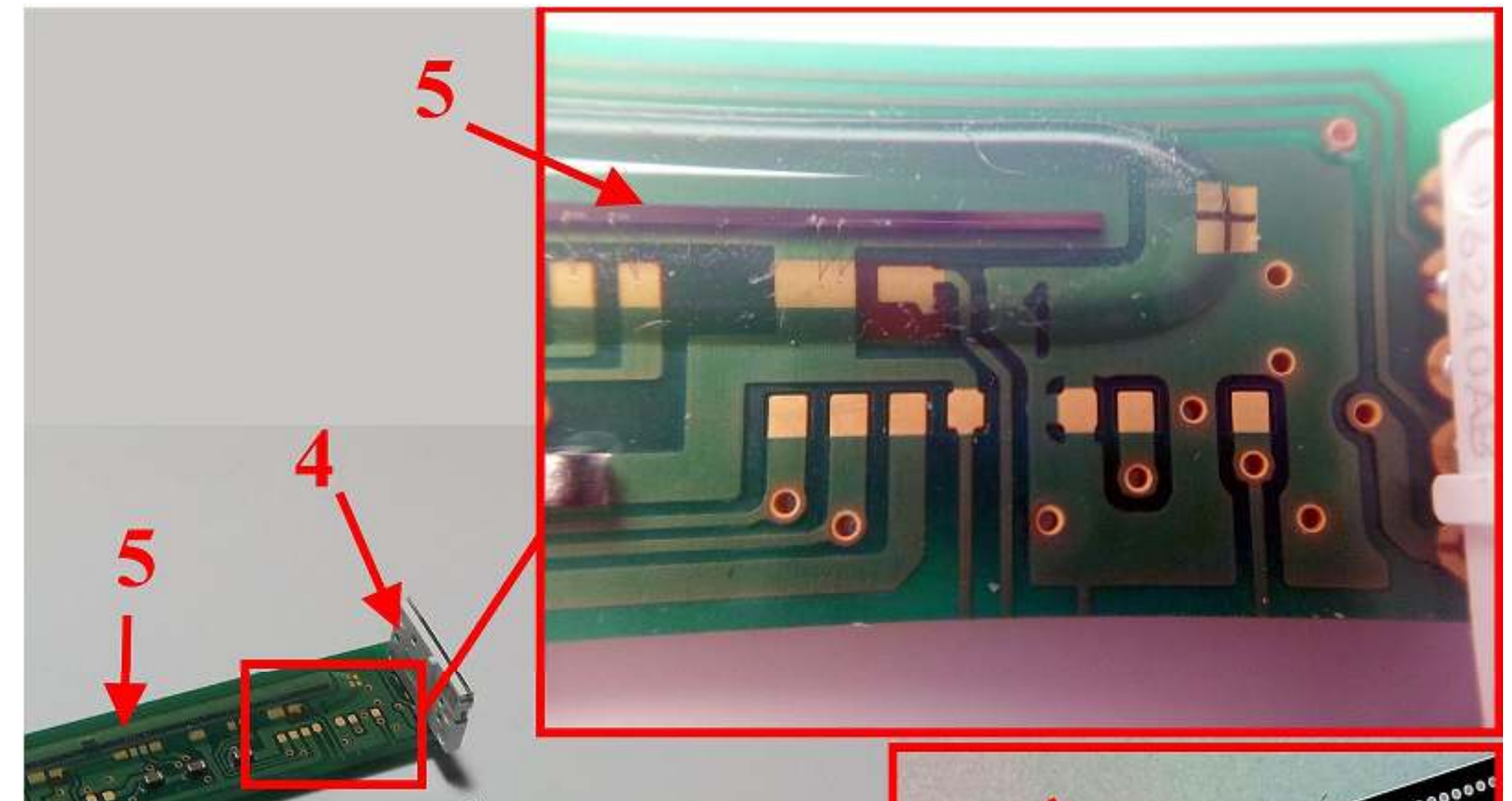
Flat bed scanners have a sensor with photodiodes arranged in lines.

Current flat bed scanners tend to have a CMOS-based CIS (*contact image sensor*):

- Cheaper
- Complex optical lense obsolete (for CCD-based scanners)
- Less power consumption – can be operated on USB power



CCD line sensor



CIS sensor and LEDs

Book Scanners



<https://www.youtube.com/watch?v=03ccxwNssmo>

<https://www.youtube.com/watch?v=cmhIJOqepVU>



Digital Images

Representing Digital Images with Three Dimensions

Images have two spatial dimensions (x,y)

Third dimension (z-axis) specifies color(s)



Precision of Discretization & Quantization

Precision of discretization: spatial resolution, e.g. amount of “dots per inch” (dpi)

Precision of quantization: color depth, i.e. number of available colors

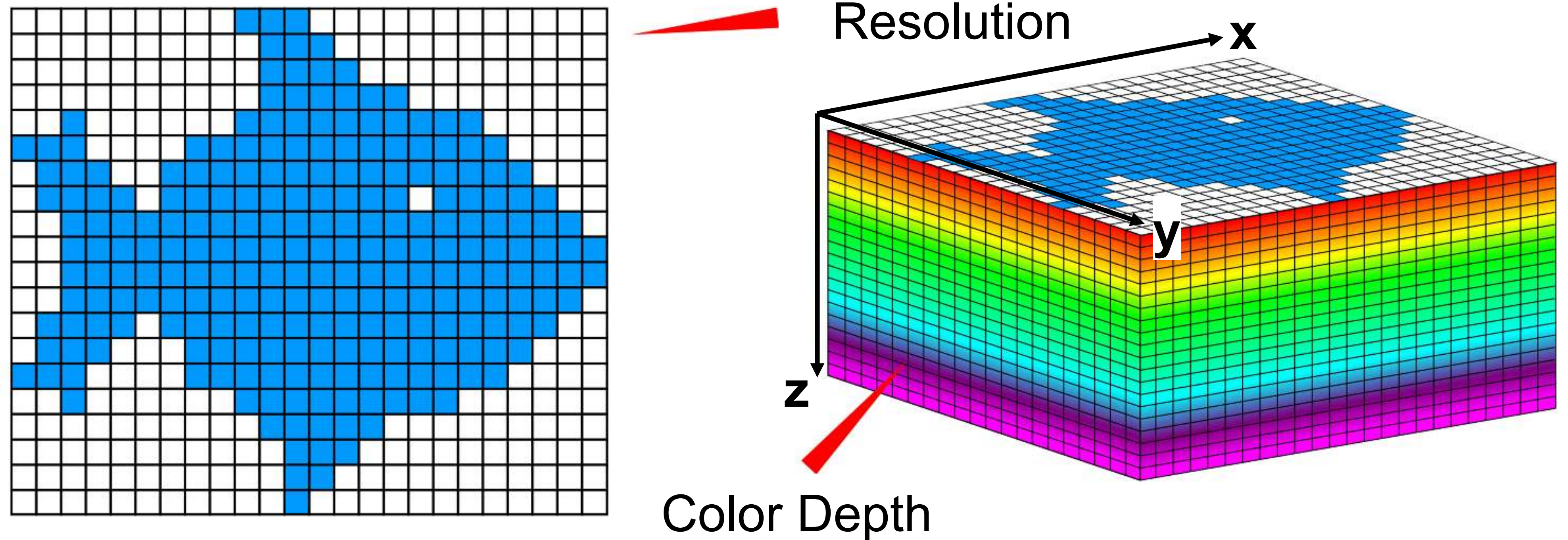


Image Parameters

Content of fish.jpg (simplified):

(0,0, 1)
(0,1, 1)
(0,2, 1)
...
(0,6, 0)
(0,7, 1)
...

x-coordinate
from left to right

(0,6, 0)

y-coordinate
from **top** to **bottom**

z-value
representing color

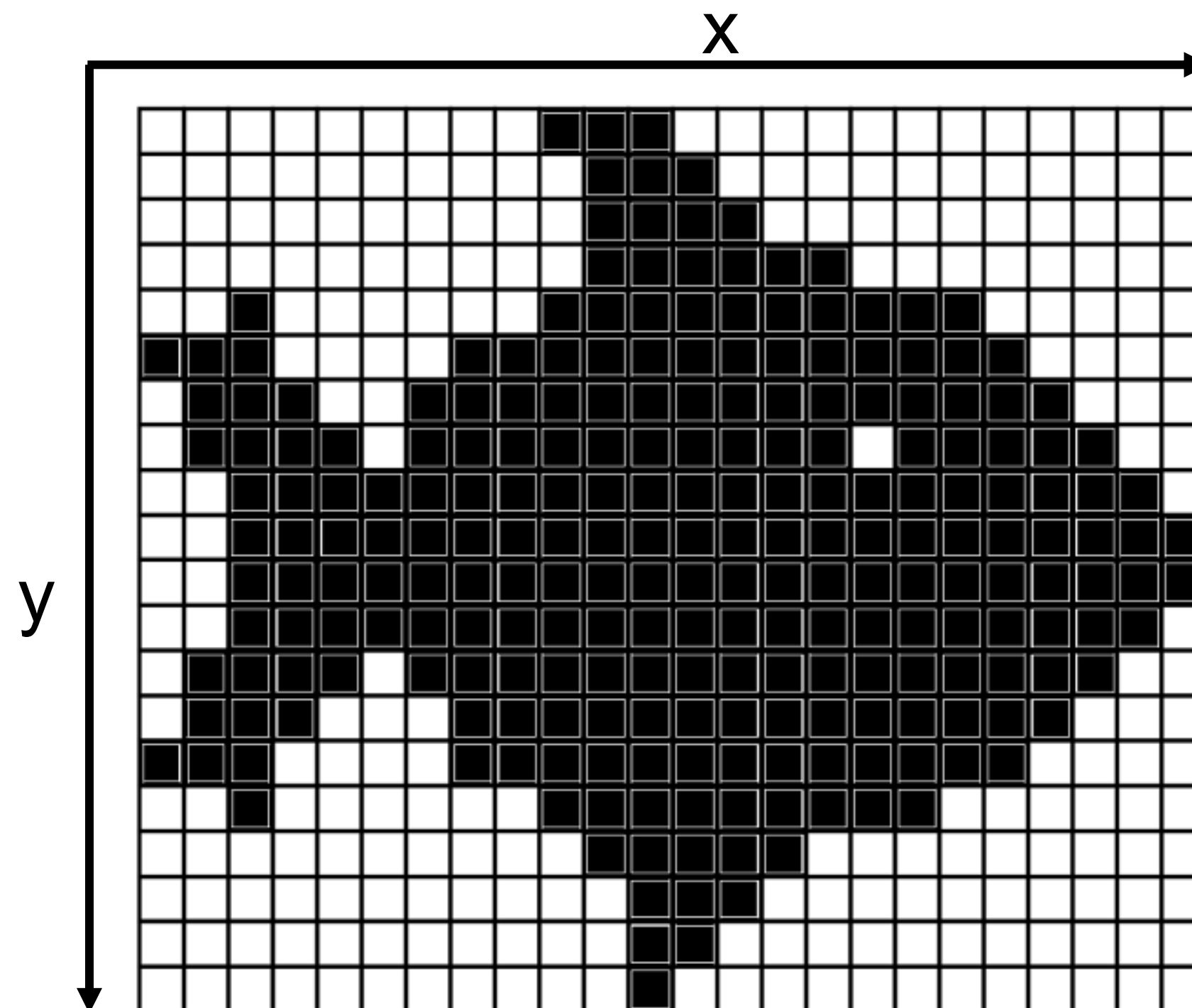


Image Parameters:

- A *pixel* is the smallest unit of an image. Size of pixel depends on output device.
- Pixel aspect ratio (not necessarily 1)
- Image size in number of pixels (e.g. 24x24 pixels)
- Image resolution: number of pixels available on a distance
 - *pixels per inch (ppi)*
 - 1 in = 2.54 cm
 - Standard resolution for displays: 72ppi
 - Typical value for print: 300ppi
- Dimension /Resolution / Pixel size:
width [px] = width [in] * resolution [ppi]

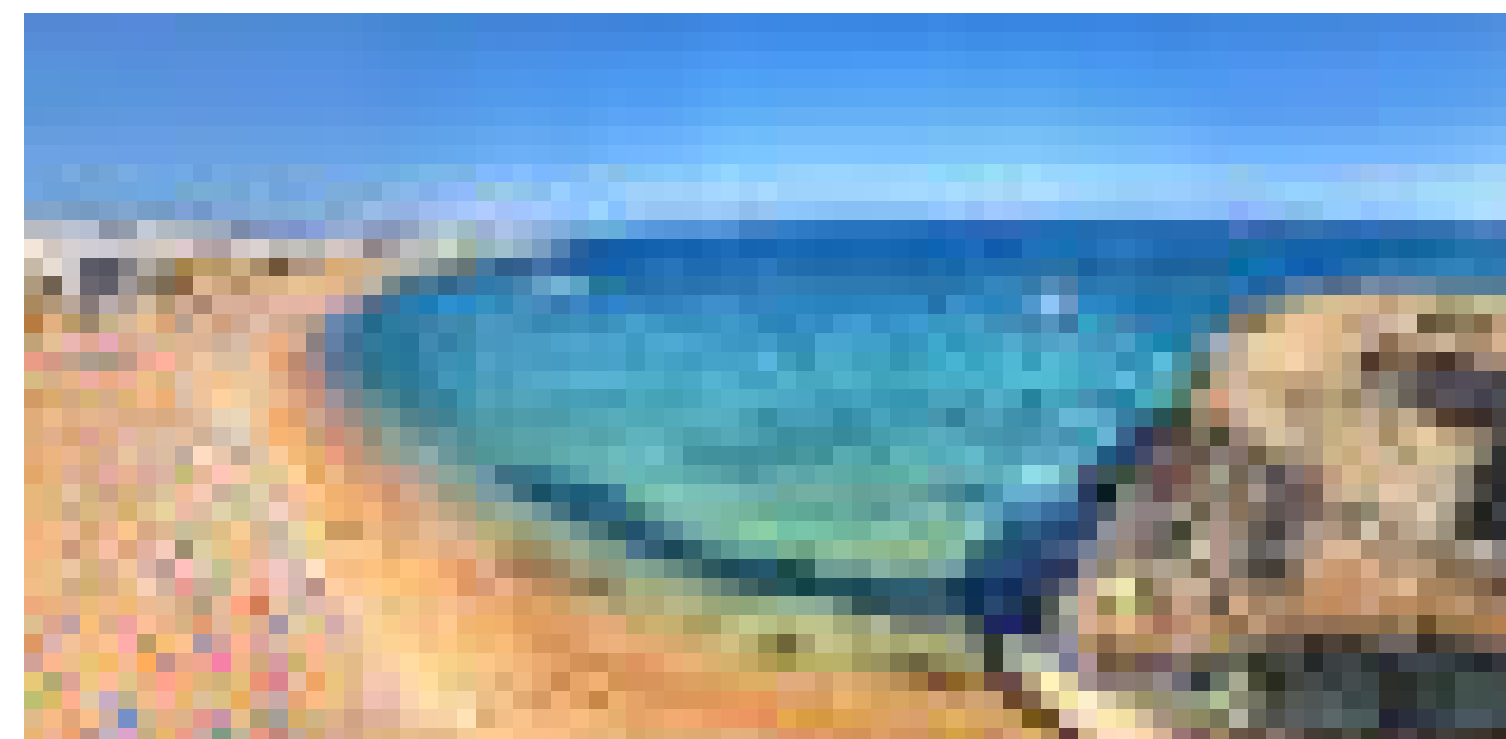
Resolution of Digital Images



320px



160px



80px



x-resolution: 1280 px

Color of Digital Images

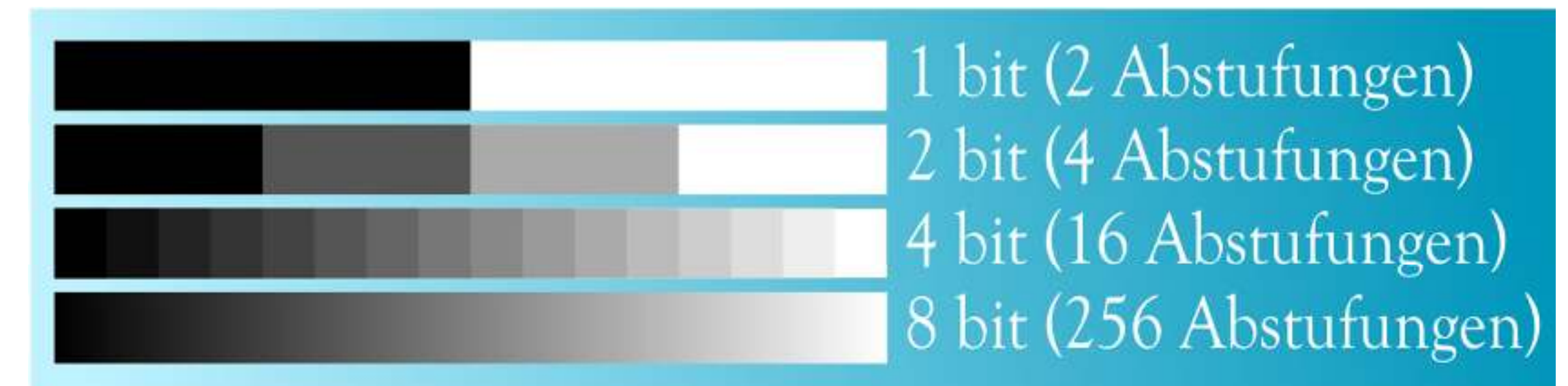
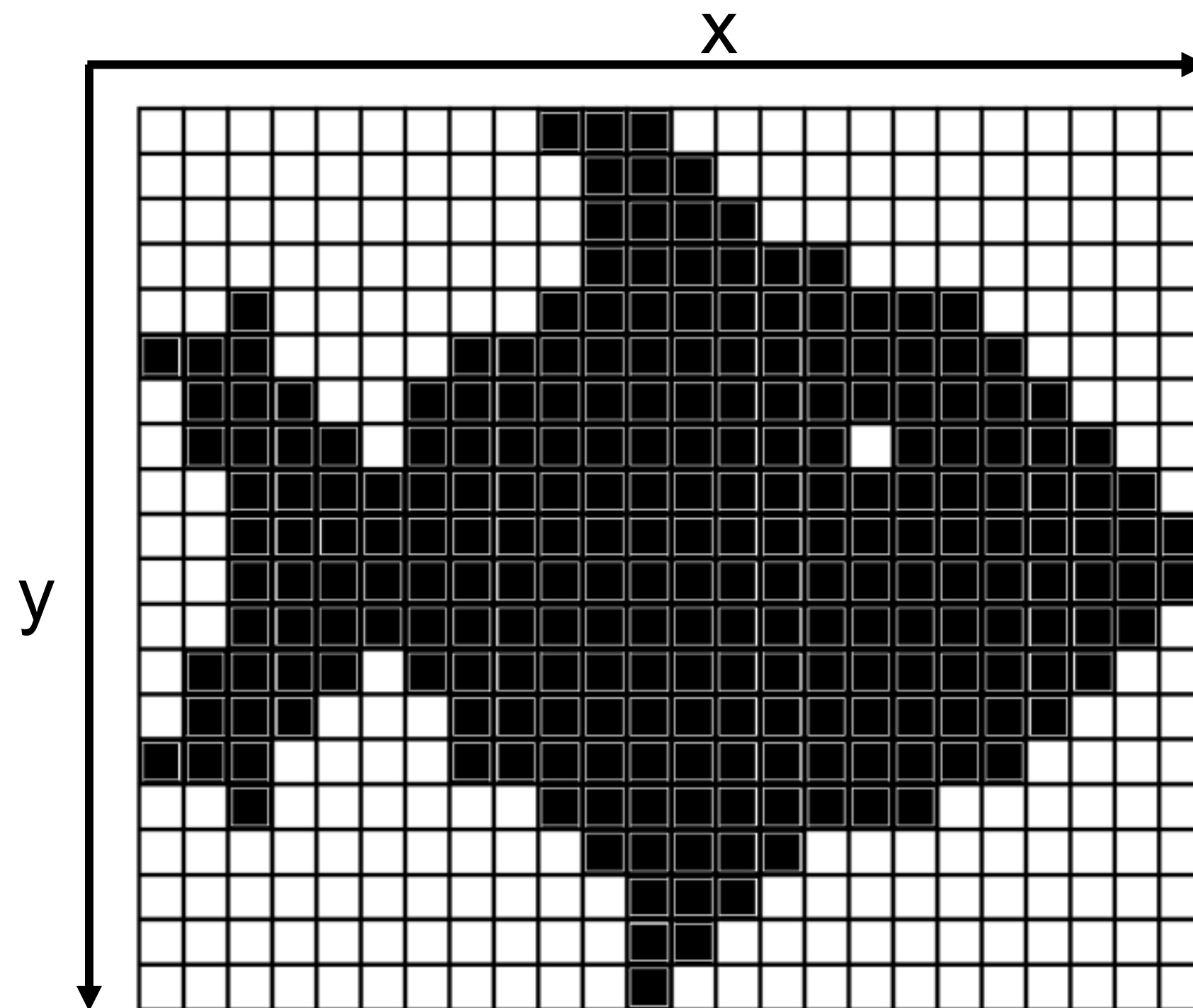
Content of fish.jpg (simplified):

(0,0, 1)
(0,1, 1)
(0,2, 1)
...
(0,6, 0)
(0,7, 1)
...

(0,6, 0)

Color Depth:

How many values are allowed here?



Color of Digital Images

Content of fish.jpg (simplified):

(0,0, 255,255,255)

(0,1, 255,255,255)

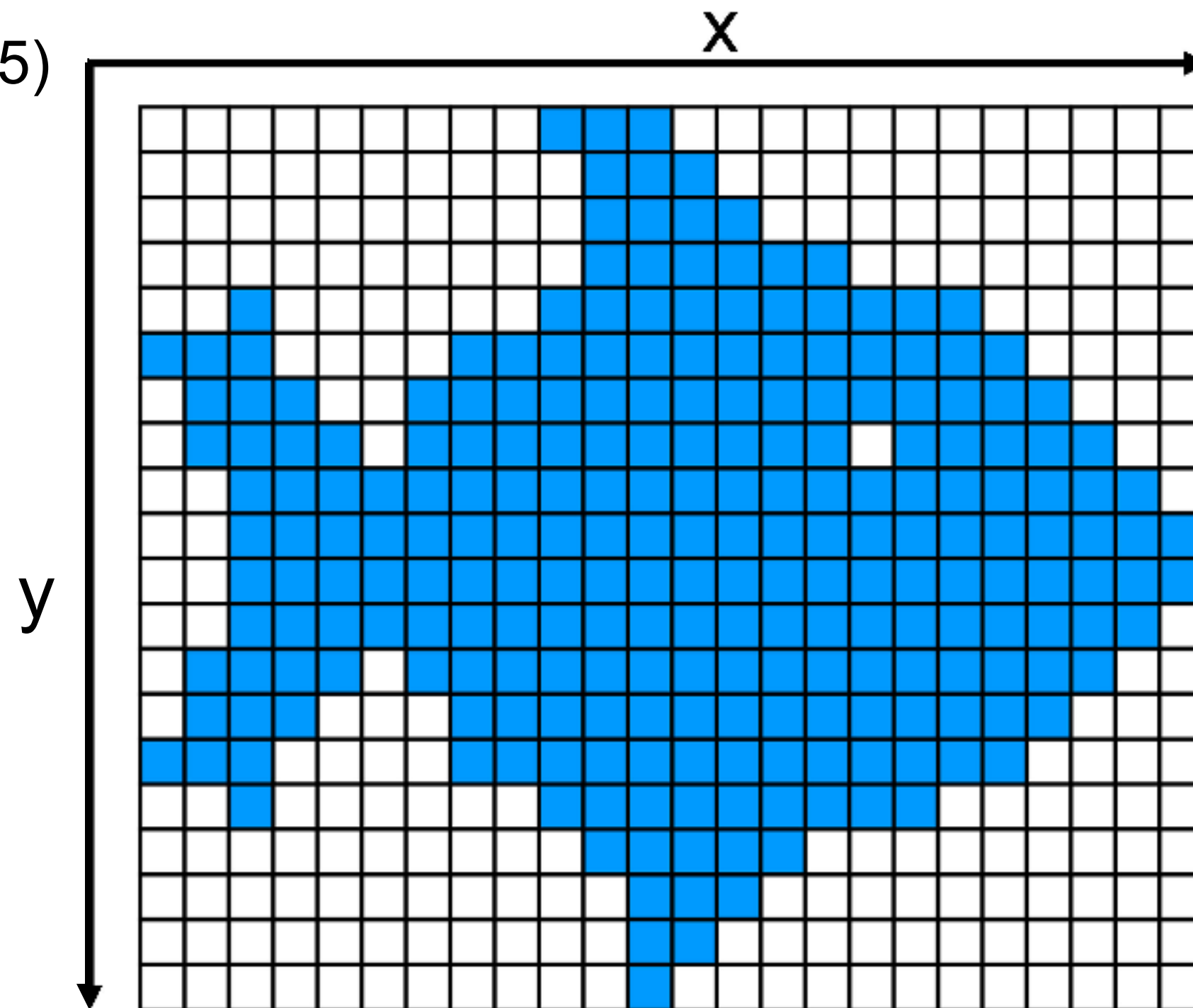
(0,2, 255,255,255)

...

(0,6, 0,153,254)

(0,7, 255,255,255)

...



(0,6, 0,153,254)

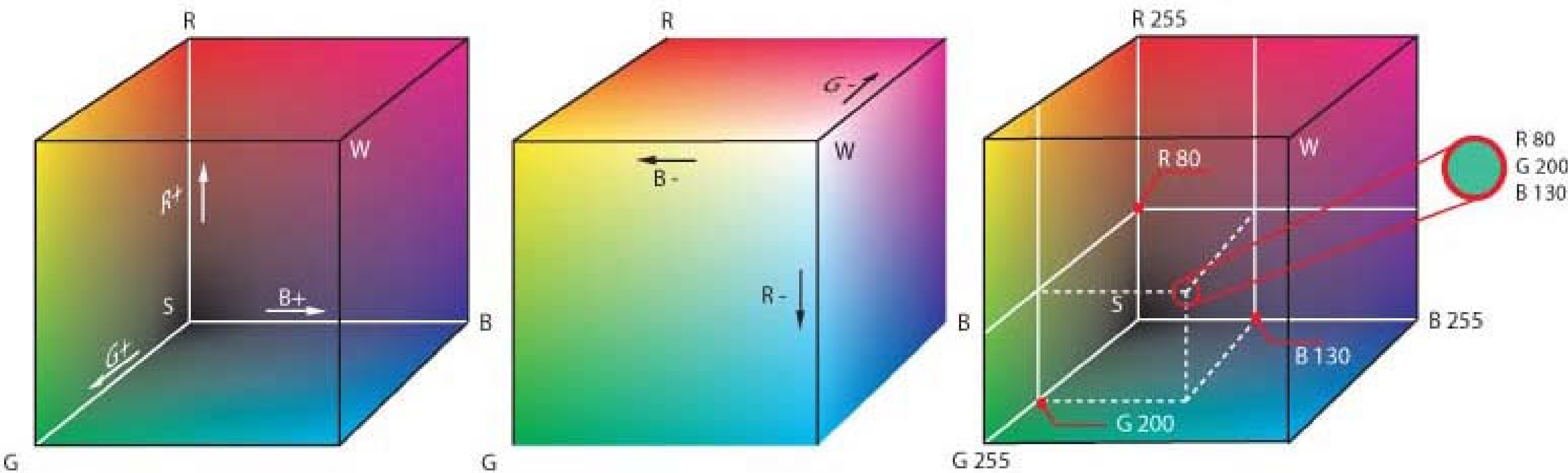
RGB color tuple

Red

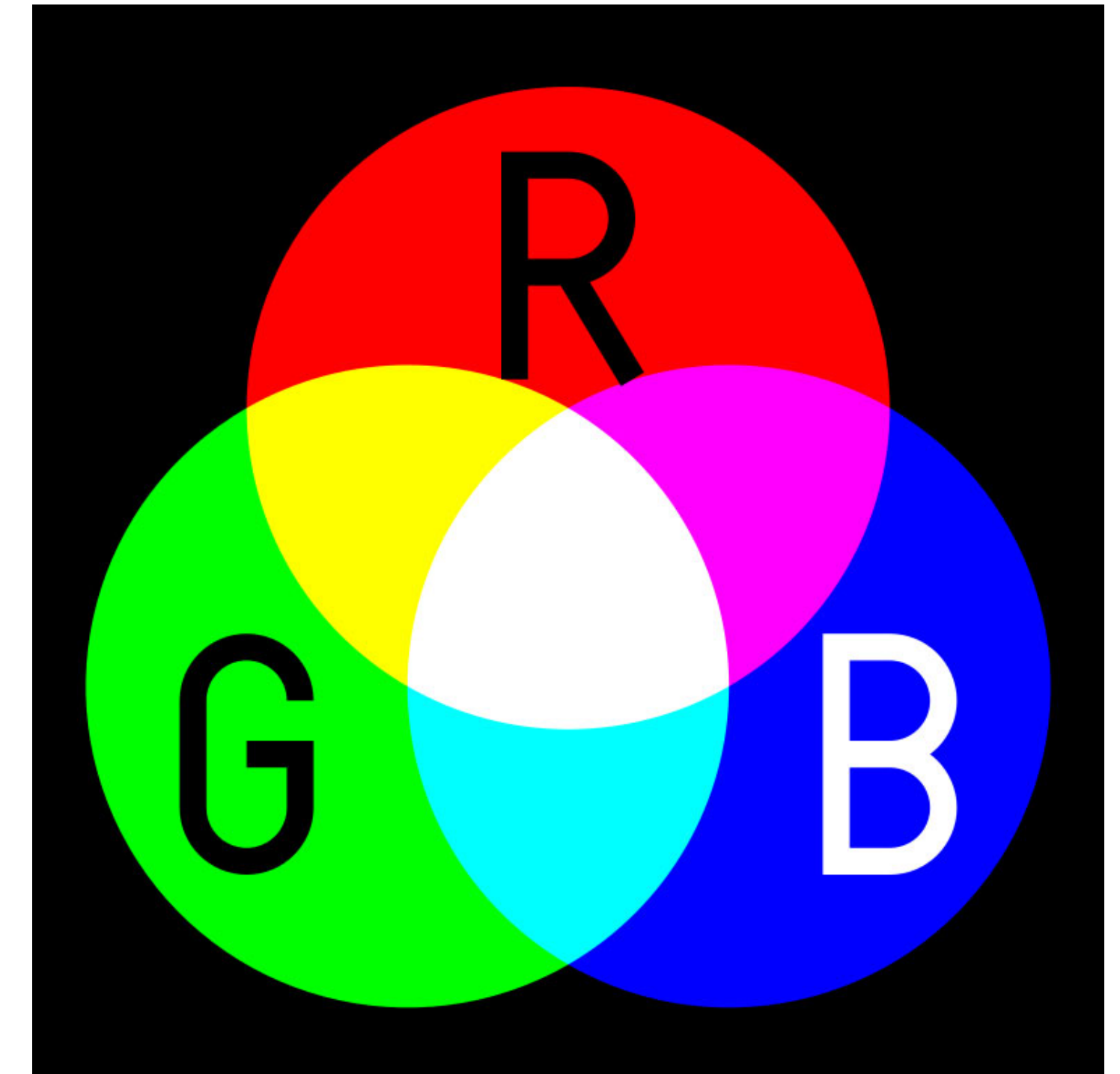
Green

Blue

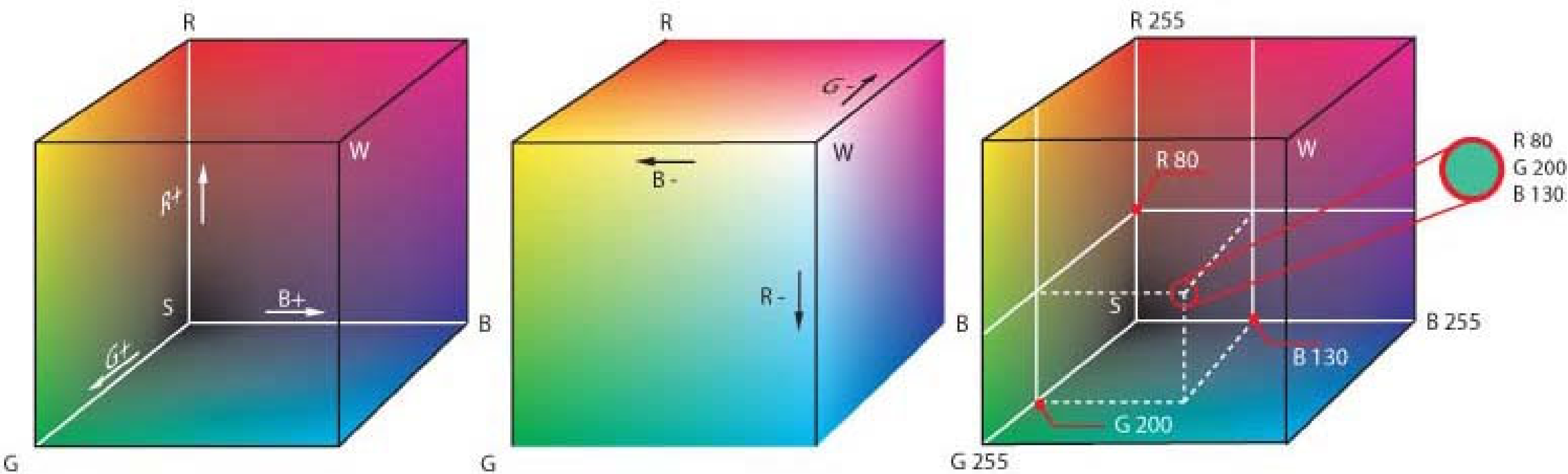
RGB Color Space



- Additive Color Model
Spectral intensities are added together to produce a color
- For output mediums that actively produce light (e.g. displays)
- Device-dependent
- Some colors not representable within the RGB space



RGB Color Space



Red
8 Bit
256 Values



Green
8 Bit
256 Values



Blue
8 Bit
256 Values

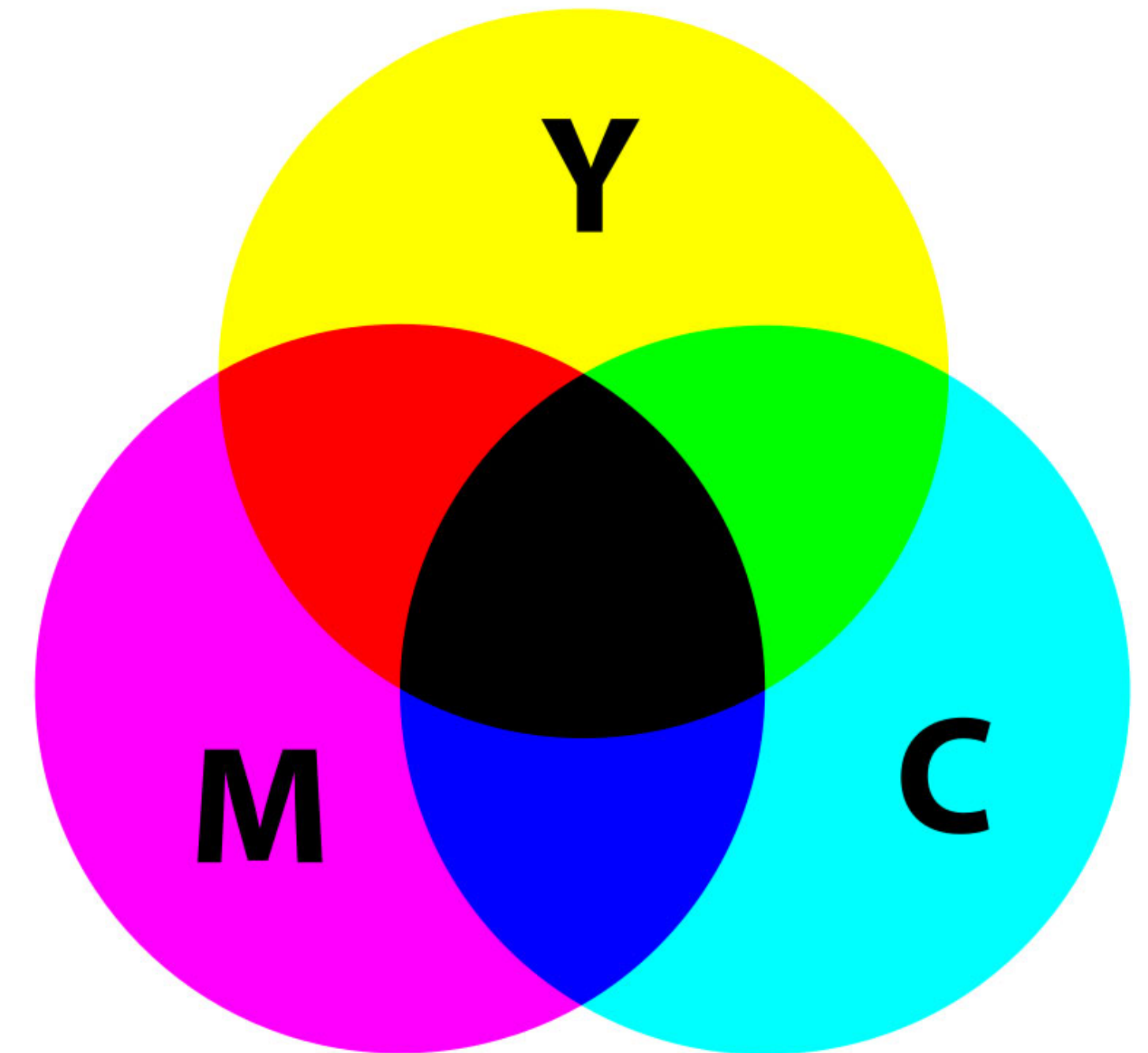


24 Red
57 Green
102 Blue

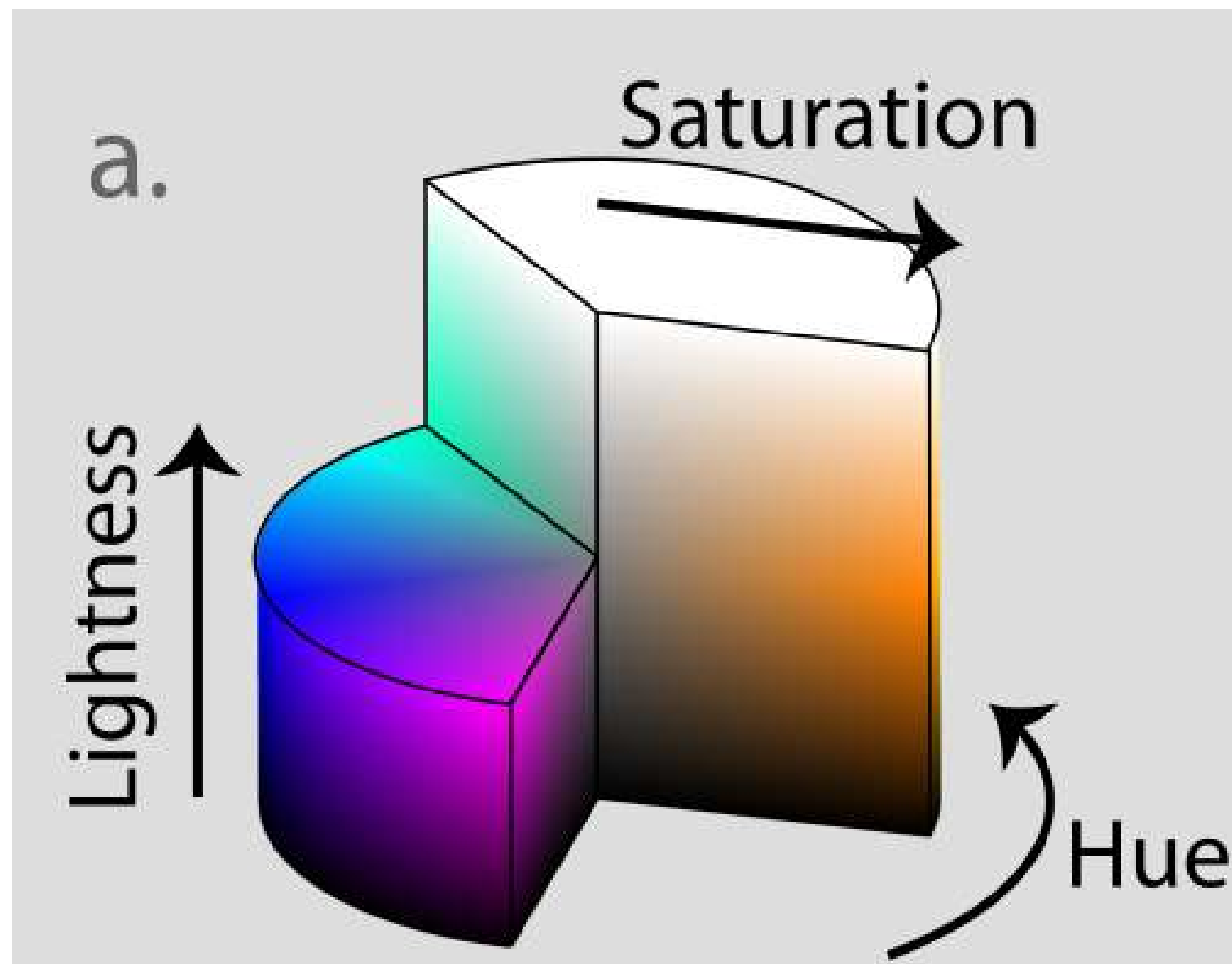
CMYK Color Space



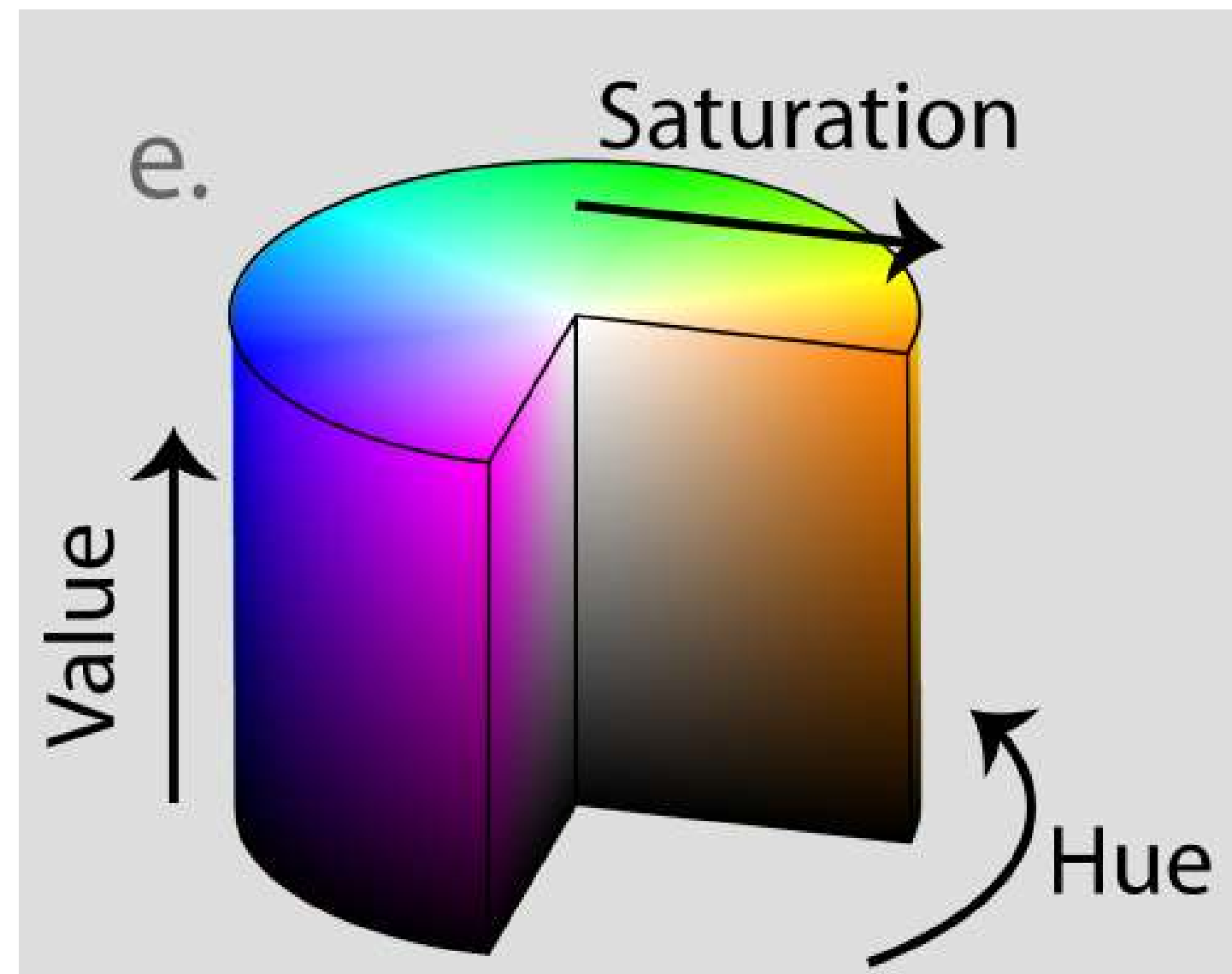
- Subtractive Color Model:
White is the natural color of the paper and color “reduces whiteness”
- Mainly used for printing
- Think of color filters
- For printing, **K** represents a black component – mainly to save ink



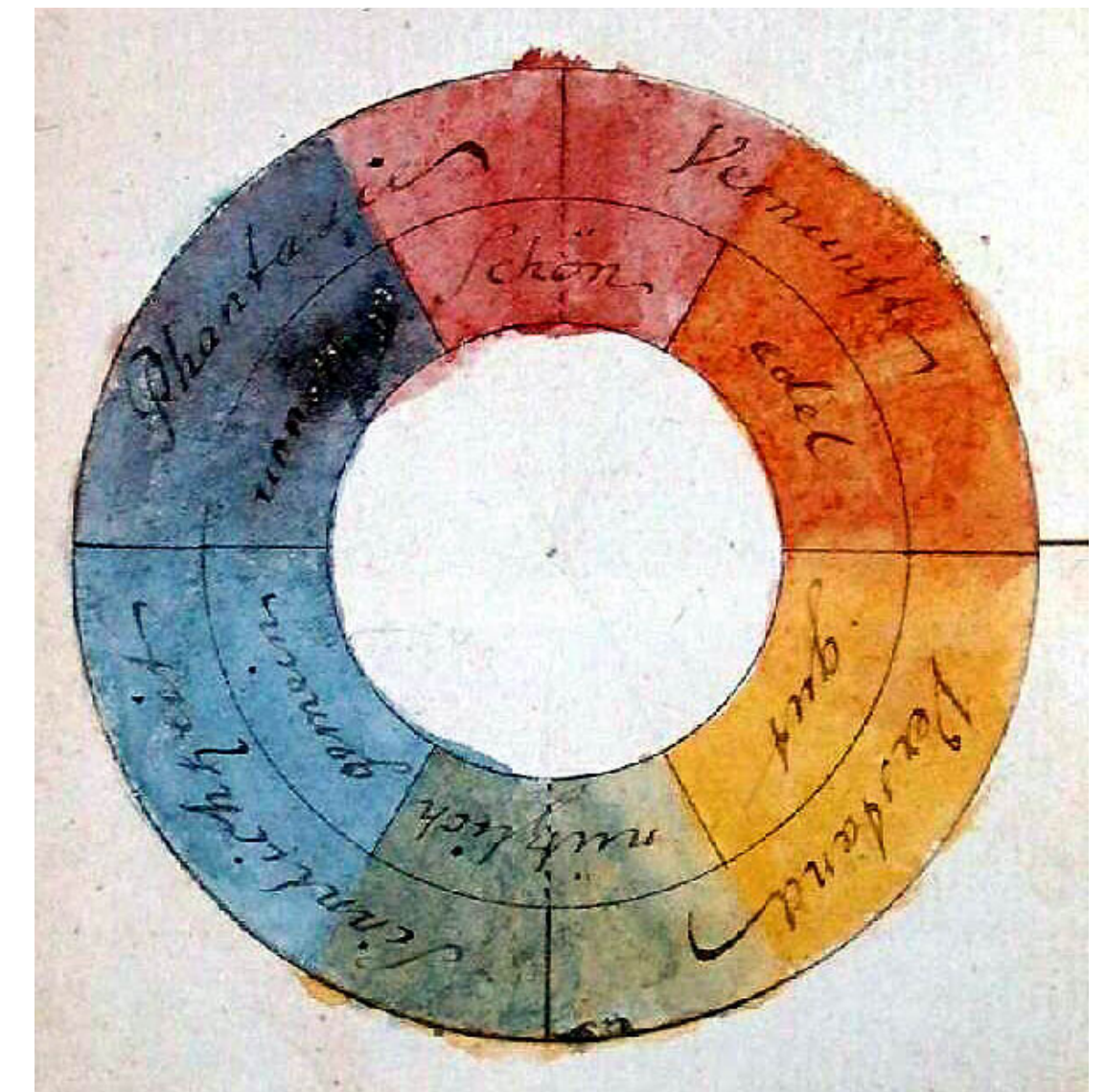
Other Color Spaces



HSL:
Hue, Saturation, Lightness

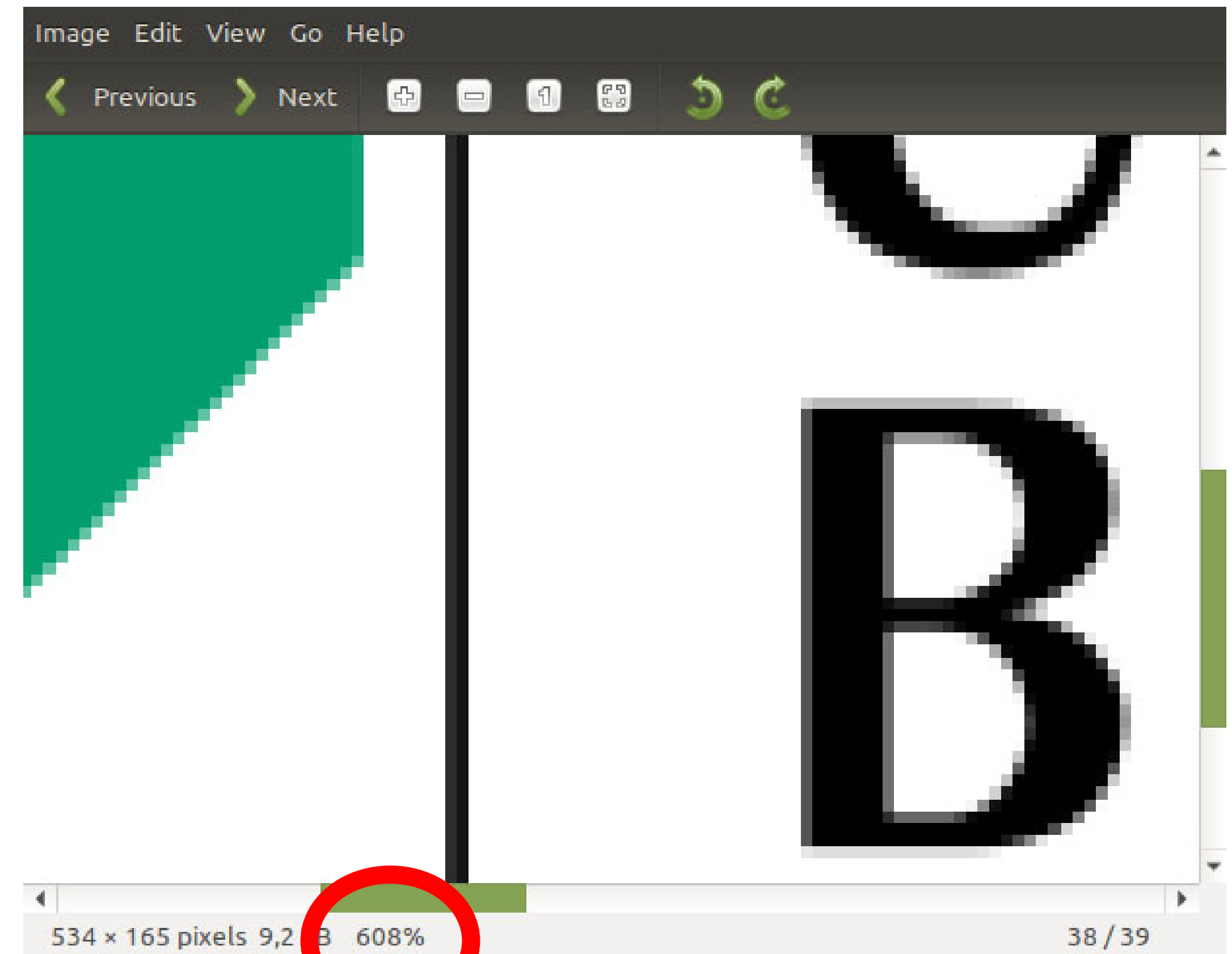
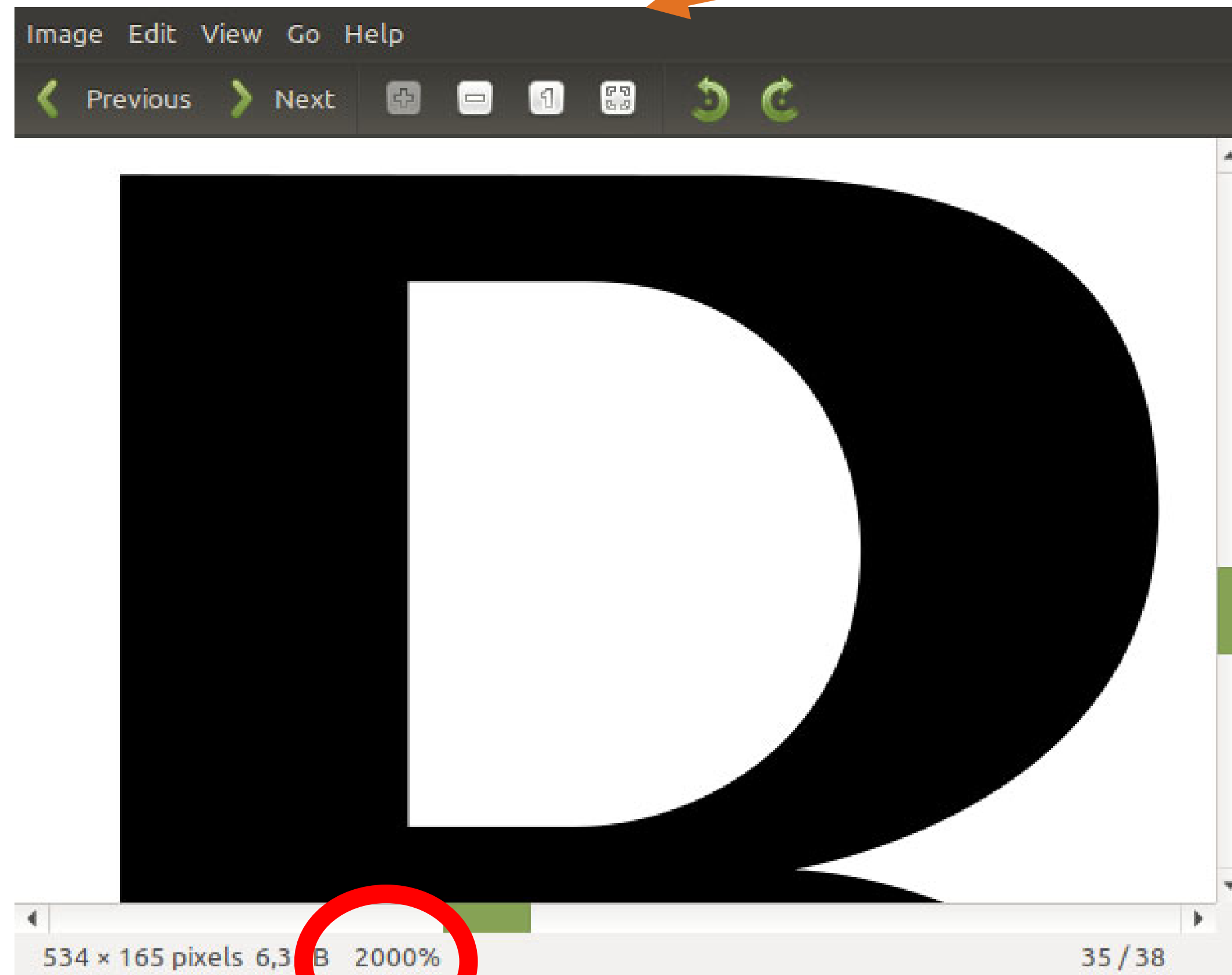


HSV:
Hue, Saturation, Value

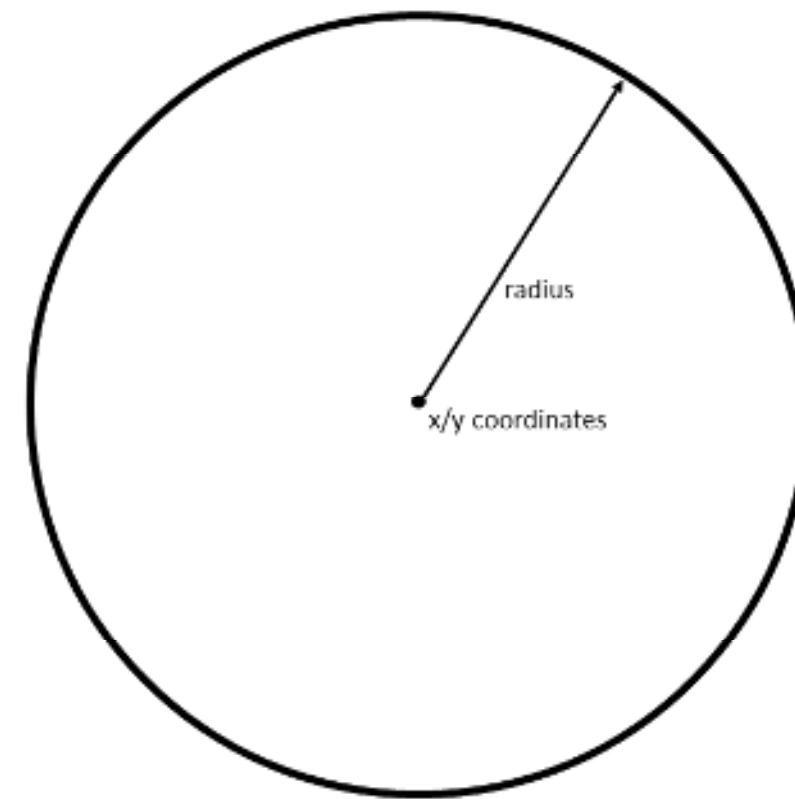


Goethe's Color Wheel

Raster vs. Vector Graphics



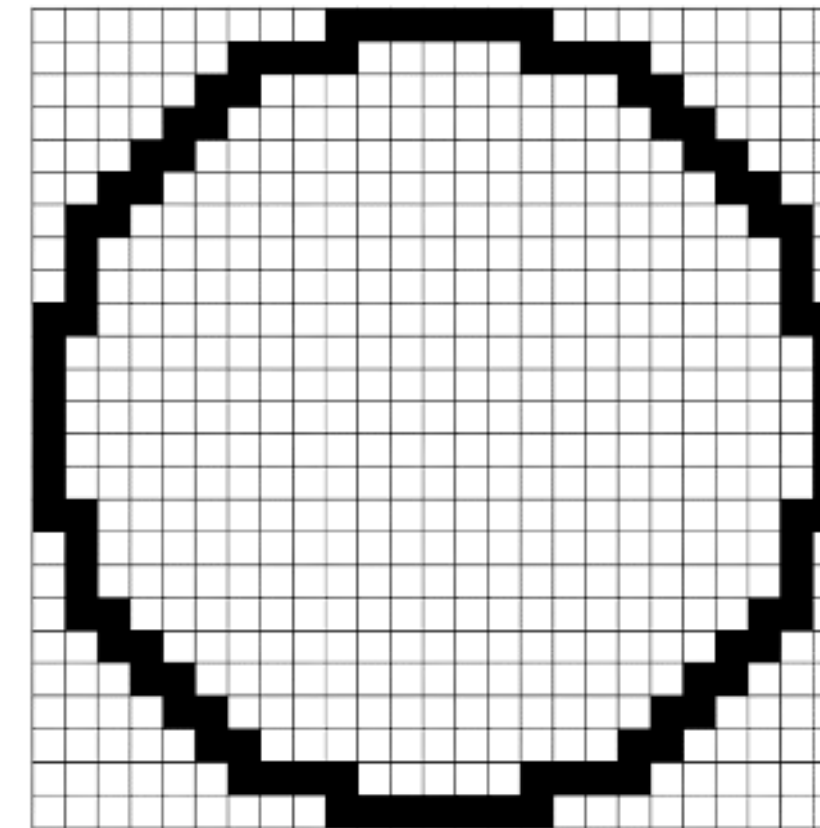
Raster vs. Vector Graphics



Vector Graphics:

Image is stored as geometric description
Scales infinitely
Small files

Great for web images, illustrations, logos, technical drawings



Raster Graphics:

Image is stored as a matrix of pixels - max. resolution fixed
Does not scale
Large files

Great for photographs

SVG: Scalable Vector Graphics

```
<svg viewBox="0 0 100 100" xmlns="http://www.w3.org/2000/svg">  
  <circle cx="50" cy="50" r="50"/>  
</svg>
```

Example SVG file

Digital Text

Text on Images

If text is digitized using a scanner, it ends up as a raster graphic.

Problem:

For the computer, this is only a matrix of pixels, e.g. color values

It is not processable as text, e.g.:

- it can't be indexed, i.e. used for search & discovery
- it can't be used for tasks of Natural Language Processing (e.g. counting words, describing the vocabulary, ...)
- it takes up much more disk space for archiving than a representation as text



Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

https://en.wikipedia.org/wiki/Optical_character_recognition

OCR: Two basic types

1. Matrix Matching

aka pattern matching or pattern recognition

Comparing an image of a character to stored glyphs pixel-by-pixel

Works best with typewritten text:

- Requires glyphs to be clearly isolated

- Stored glyphs need to be in a similar font and in a similar scale

OCR: Two basic types

2. Feature Recognition

based on machine-learning:

Decompose glyphs into “features” like lines, closed loops, line direction, and line intersections

Describe glyphs by abstract representations, i.e. numeric vectors that express the features mathematically

Compare features to abstract representations of glyphs by comparing vectors, e.g. find the most similar match

Helping the computer: two crucial steps before OCR

1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)

2. Segmentation

1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation

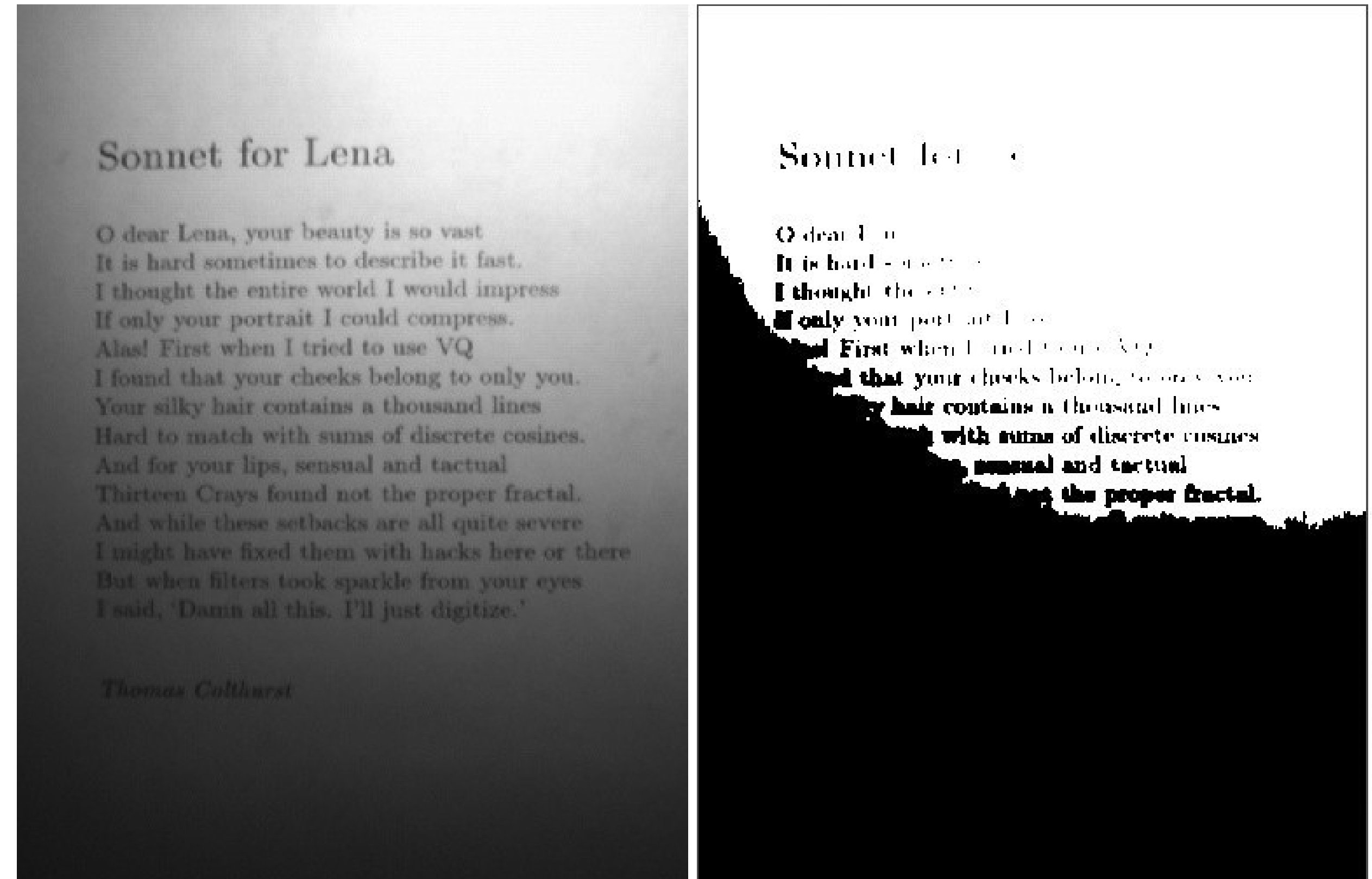
Helping the computer: two crucial steps before OCR

1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)

2. Segmentation

1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation



Helping the computer: two crucial steps before OCR

1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)

2. Segmentation

1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation

The Energy Picture: Where Are We Now? Where Are We Headed?
EPA's experience, through its interactions with U.S. companies, is that many are initiating energy programs. For companies operating formal energy programs, these programs are typically less than 5 years old. And, the involvement of senior executives in energy planning and decision-making is just beginning.

Market trends suggest that the demand for energy resources will rise dramatically over the next 25 years:

Global demand for all energy sources is forecast to grow by 57% over the next 25 years.

U.S. demand for all types of energy is expected to increase by 31% within 25 years.

By 2030, 56% of the world's energy use will be in Asia.

Electricity demand in the U.S. will grow by at least 40% by 2032.

New power generation equal to nearly 300 (1,000MW) power plants will be needed to meet electricity demand by 2030.

Currently, 50% of U.S. electrical generation relies on coal, a fossil fuel; while 85% of U.S. greenhouse gas emissions result from energy-consuming activities supported by fossil fuels.

Sources: Annual Energy Outlook (DOE/EIA-0383(2007)), International Energy Outlook 2007 (DOE/EIA-0484(2007)), Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2005 (April 2007) (EPA 430-R-07-002)

If energy prices also rise dramatically due to increased demand and constrained supply business impacts could include:

The Energy Picture: Where Are We Now? Where Are We Headed?
EPA's experience, through its interactions with U.S. companies, is that many are initiating energy programs. For companies operating formal energy programs, these programs are typically less than 5 years old. And, the involvement of senior executives in energy planning and decision-making is just beginning.

Market trends suggest that the demand for energy resources will rise dramatically over the next 25 years:

Global demand for all energy sources is forecast to grow by 57% over the next 25 years.

U.S. demand for all types of energy is expected to increase by 31% within 25 years.

By 2030, 56% of the world's energy use will be in Asia.

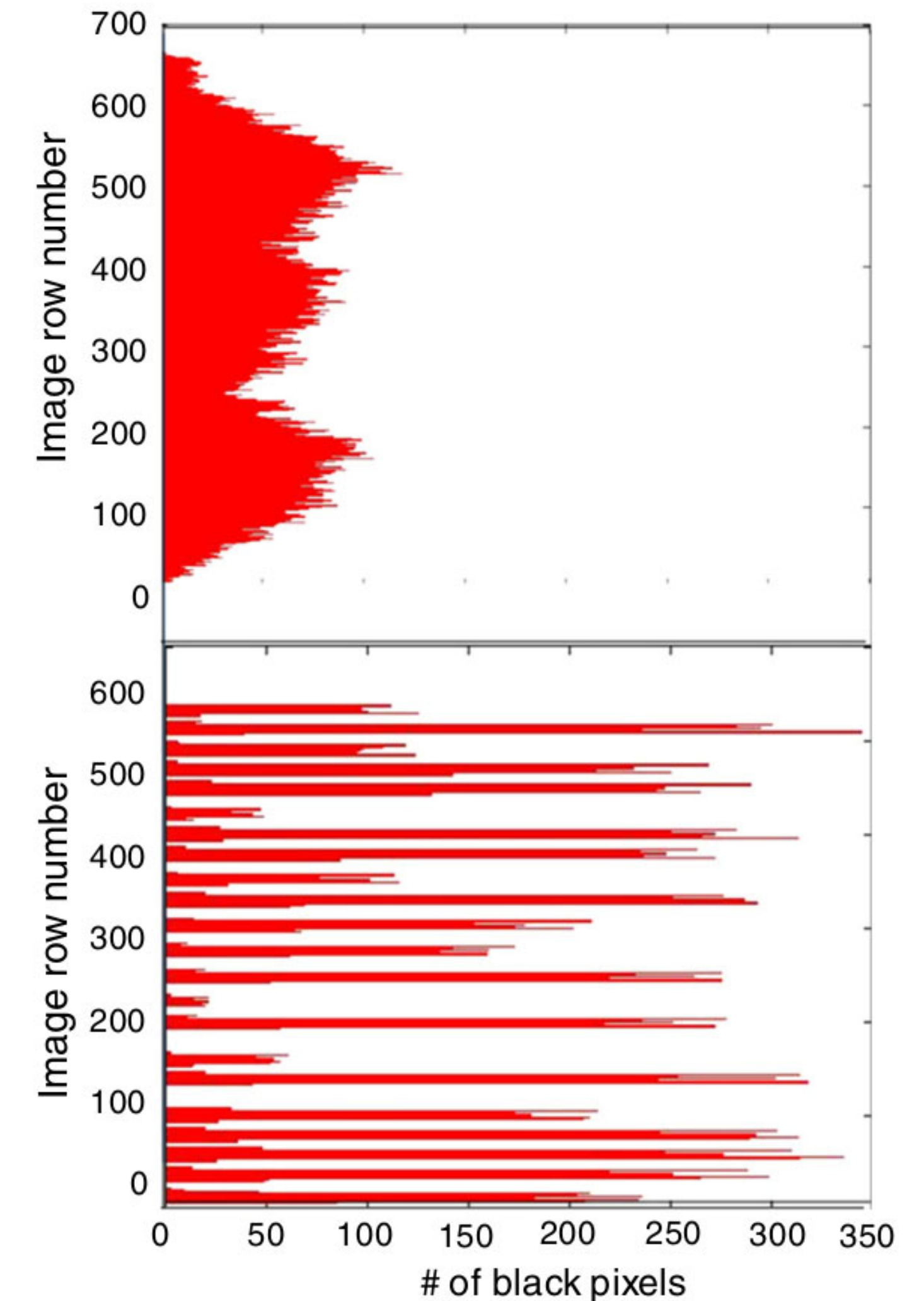
Electricity demand in the U.S. will grow by at least 40% by 2032.

New power generation equal to nearly 300 (1,000MW) power plants will be needed to meet electricity demand by 2030.

Currently, 50% of U.S. electrical generation relies on coal, a fossil fuel; while 85% of U.S. greenhouse gas emissions result from energy-consuming activities supported by fossil fuels.

Sources: Annual Energy Outlook (DOE/EIA-0383(2007)), International Energy Outlook 2007 (DOE/EIA-0484(2007)), Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2005 (April 2007) (EPA 430-R-07-002)

If energy prices also rise dramatically due to increased demand and constrained supply business impacts could include:



Helping the computer: two crucial steps before OCR

1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)

2. Segmentation

1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation



Helping the computer: two crucial steps before OCR

1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)

2. Segmentation

1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation



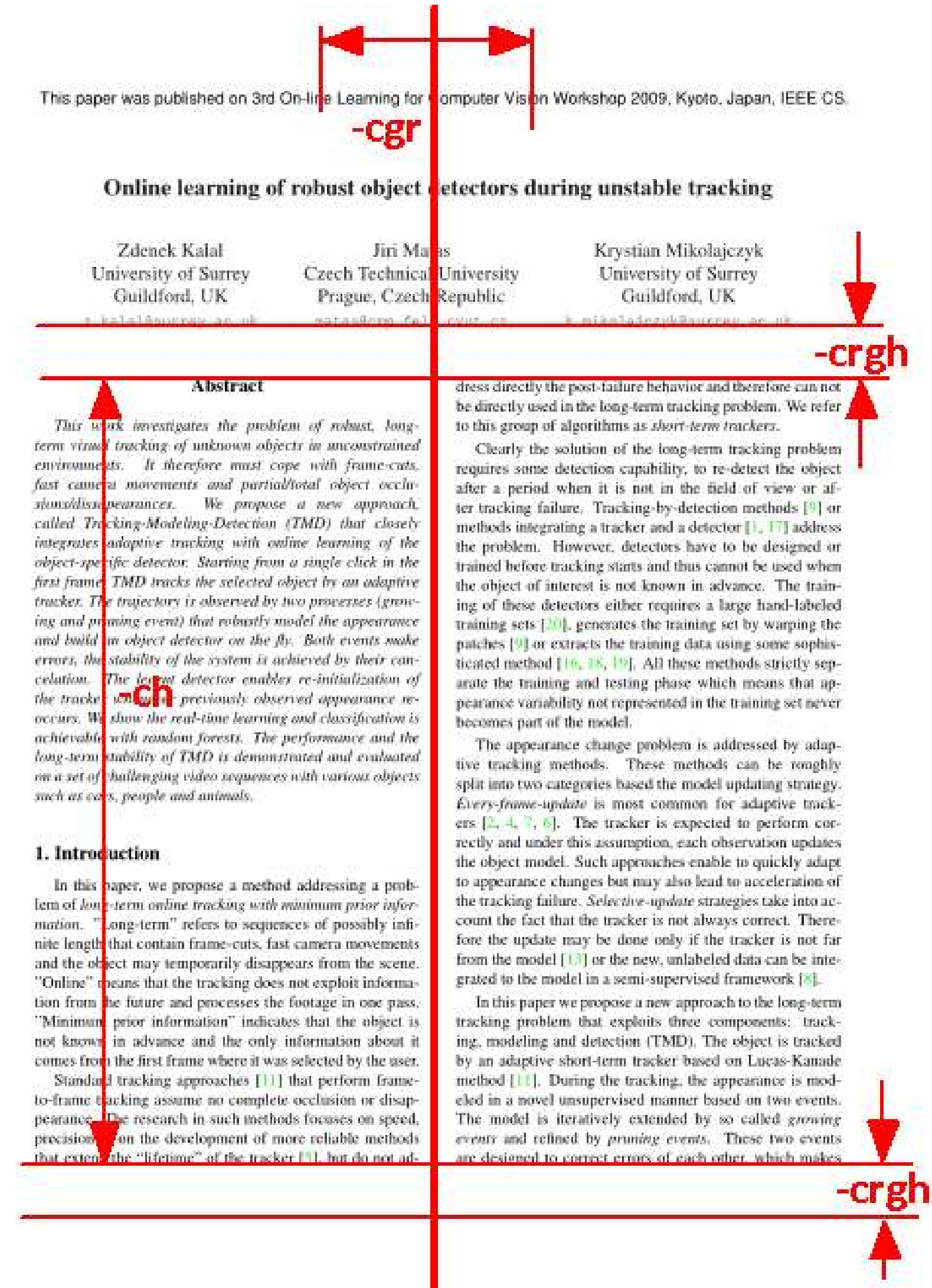
Helping the computer: two crucial steps before OCR

1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)

2. Segmentation

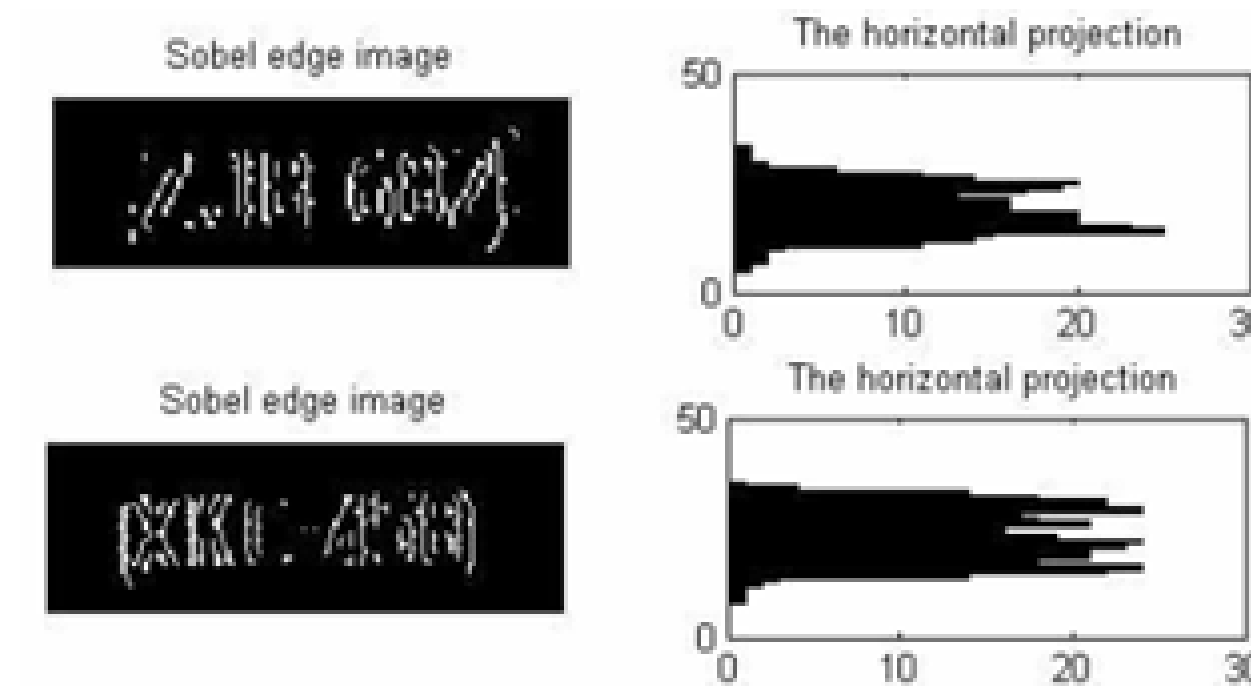
1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation



Helping the computer: two crucial steps before OCR

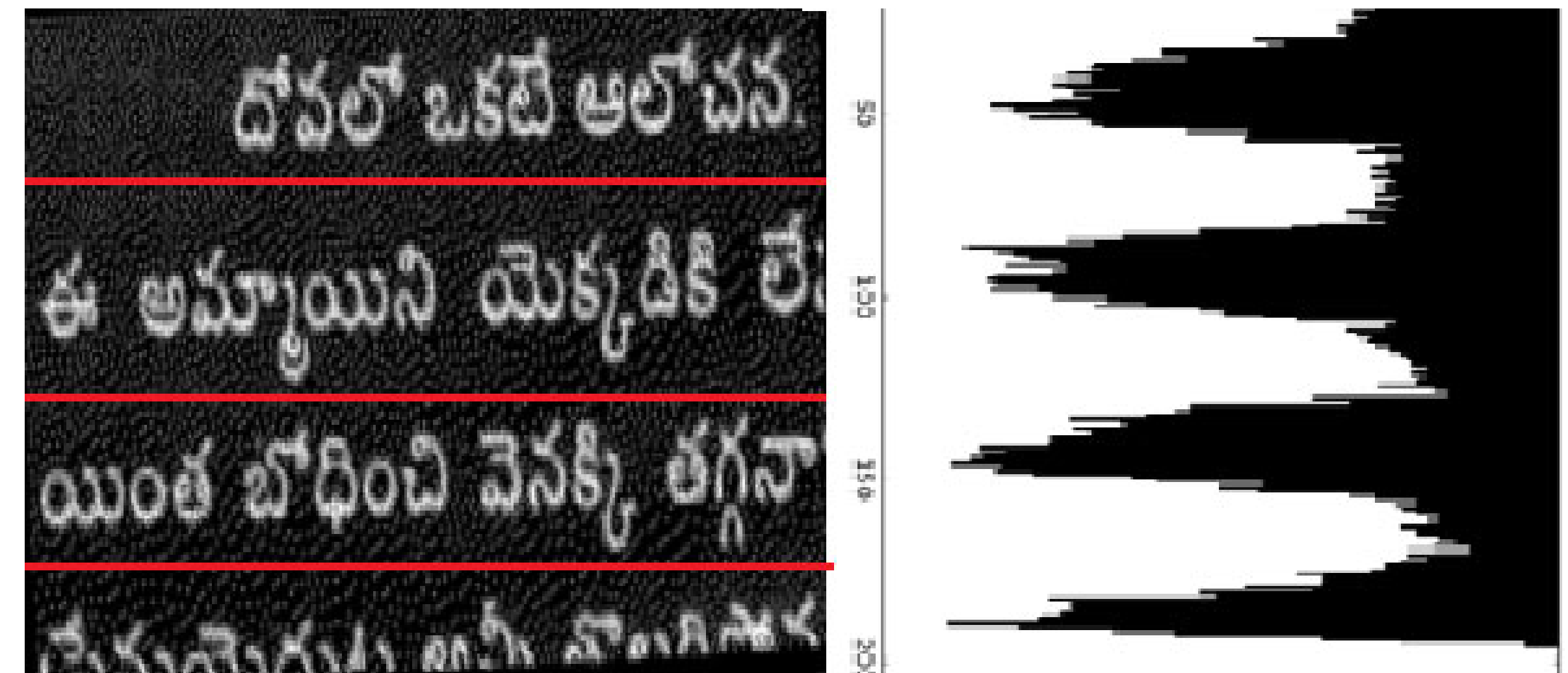
1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)



2. Segmentation

1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation



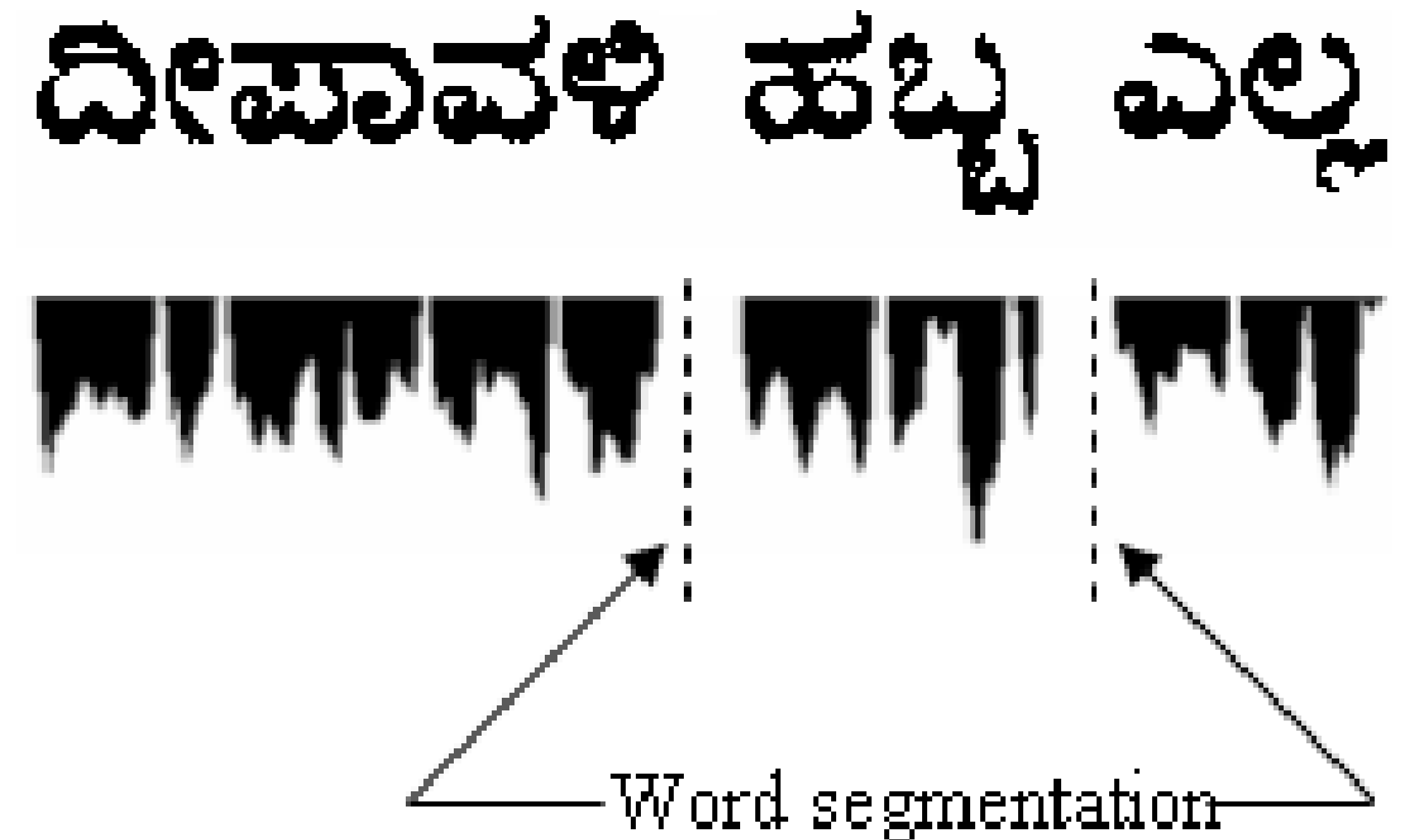
Helping the computer: two crucial steps before OCR

1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)

2. Segmentation

1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation



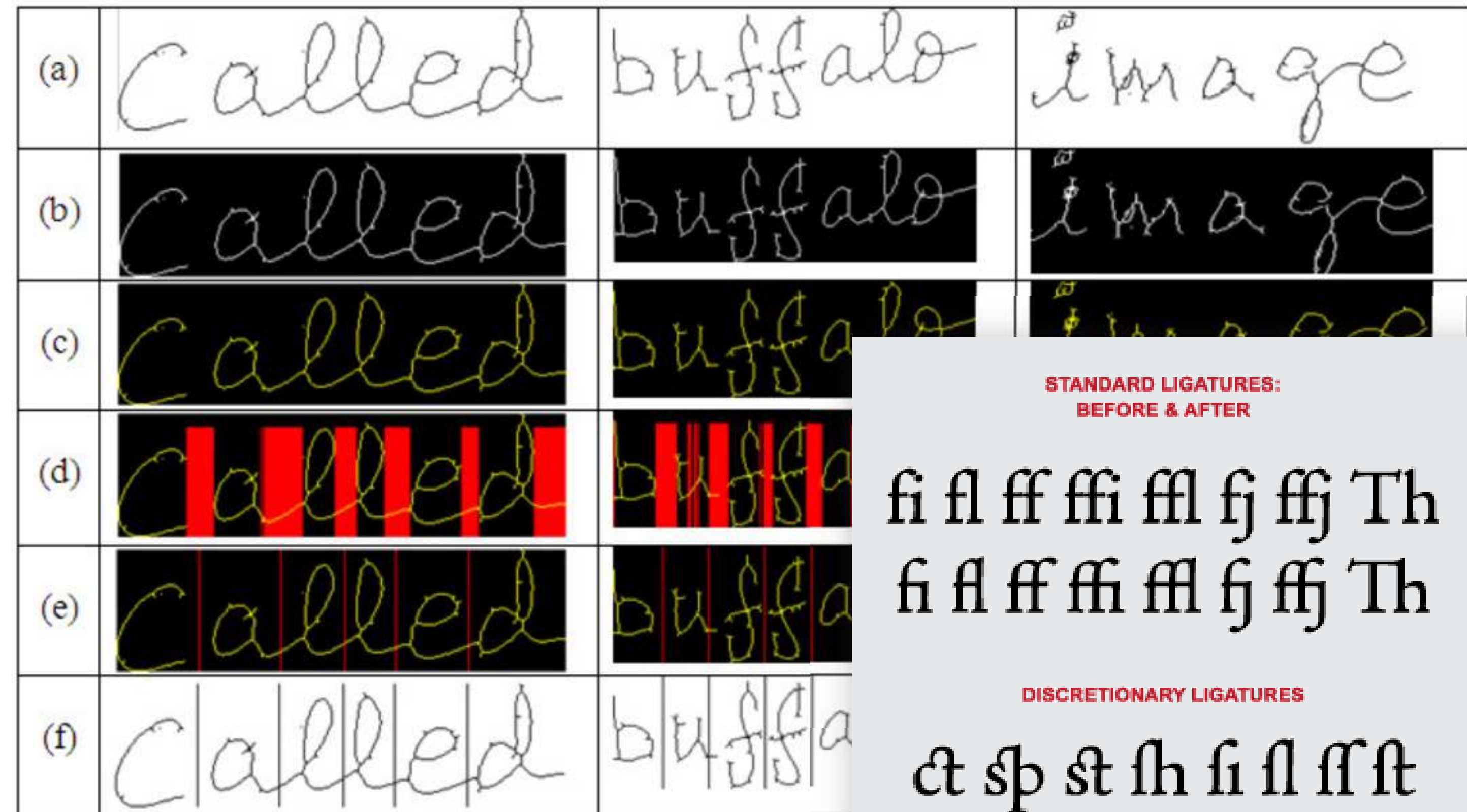
Helping the computer: two crucial steps before OCR

1. Preprocessing

1. Binarization
2. Skew Correction
3. Noise Removal
4. Thinning and Skeletonization
(only for handwritten text)

2. Segmentation

1. Layout-Analysis
2. Line-level Segmentation
3. Word-level Segmentation
4. Character-level Segmentation



**STANDARD LIGATURES:
BEFORE & AFTER**

fi fl ff ffi ffl fj ffj Th
fi fl ff ffi ffl fj ffj Th

DISCRETIONARY LIGATURES

ct sp st th fi fl ff ft
ca la nt ra st th o ss
pe fr sk off

Helping the computer: what to do after OCR

- Dictionary-based approach

Replace words by most likely words in a dictionary

Decision of requiring a replacement is based on a distance metric between words, e.g.

Damerau-Levenshtein distance, an edit distance expressing the number of operations (insertions, deletions, substitutions of characters) required to turn one word into another

Medical Mÿstory  Medical Mystery

- Context-based approach

Based on Statistical Language Modeling, e.g. modeling the probabilities of sequences of characters

Statistical distributions of word associations in character sequences

Takes context into account

Medical Mÿstory  Medical History

OCR Accuracy

Character Accuracy	Approach	Purpose
99,997%	Double keying with corrections	Editions
99,95%	Double keying	Negative Search
99%	OCR with near-perfect input	Text Mining
95%	OCR with good input	Positive Search

Thanks.

mirco.schoenfeld@uni-bayreuth.de